Journal of Political Economy

Volume 82 Number 1 January/February 1974

Articles

1 Benefit-Cost Analysis and Trade Policies
Partha Dasgupta and Joseph E. Stiglitz

34 Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition
Sherwin Rosen

56 A Diagrammatic Exposition of the Harberger Model with One Immobile Factor
Charles E. McLure, Jr.

83 An Economic Theory of Suicide
Daniel S. Hamermesh and Neal M. Soss

99 John Maynard Keynes: Scientist or Politician?
Elizabeth Johnson

112 Group Segregation and Optimal Jurisdictions
Martin McGuire

133 Stability and the Quantity Theory
P. H. Howitt

Communications

152 The Interrelation between Labor and Capital Components of Racial Income Differences
George M. von Furstenberg

163 An Optimal Program for Managing Common Property Resources with Congestion Externalities
Gardner Brown, Jr.

174 Cognitive Range in the Theory of Revealed Preference
Anthony Y. C. Koo and Peter Schmidt

180 Consumers' Durable Goods, Human Capital, and the Pareto Optimality of the Lange-Lerner State
Michael R. Darby

JOURNAL OF
POLITICAL ECONOMY

Edited by
HARRY G. JOHNSON and GEORGE J. STIGLER
ROBERT BARRO, Associate Editor
In cooperation with OTHER MEMBERS of the DEPARTMENT OF
ECONOMICS and the GRADUATE SCHOOL OF BUSINESS
of the UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES
(SEE INSIDE BACK COVER)
DIANE WELCH, Editorial Assistant

Accession Numbers
40423
Date 12-5-1975

The **Journal of Political Economy** is published bimonthly in February, April, June, August, October, and December by The University of Chicago Press Subscription rates, U S A institutions, 1 year $20 00, 2 years $38 00, 3 years $55 00, individuals, 1 year $15 00, 2 years $28 00, 3 years $40 00 Student subscription rate, U S A 1 year $9 00 (letter from professor must accompany subscription) Other countries add $1 00 for each year's subscription to cover postage Single copy rates institutions $4 00, individuals $3 00 Supplements institutions $4 50, individuals $3 50 Back issues are available from 1962 (vol 70) Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent

Claims for missing numbers should be made within the month following the regular month of publication The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit

Business correspondence should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637 **Manuscripts should be submitted in duplicate (3 copies if possible), accompanied by a $10 00 submission fee made payable to the Journal** Accepted manuscripts must be typed according to the University of Chicago *Manual of Style* References should be typed double-spaced at the end of the article Footnotes should be numbered in sequence and double-spaced following the references Tables should follow the footnotes Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript

Applications for permission to quote from this journal should be addressed to The University of Chicago Press and will be freely granted Reproduction of articles requires written permission from the publisher and the author

Reprinted volumes 1-72 available from Walter J Johnson, Inc , 111 Fifth Avenue, New York, New York 10003 Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, in **microfiche** from J S Canner & Co , 49-65 Lansdowne Street, Boston, Massachusetts 02215

Notice to subscribers If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address *Allow four weeks for the change*

Second-class postage paid at Chicago, Illinois, and at additional mailing office
© 1974 by the University of Chicago All rights reserved
ISSN 0022 3808
PRINTED IN U S A

Journal of Political Economy

Volume 82 Number 1 January/February 1974

Articles

1 Benefit-Cost Analysis and Trade Policies
Partha Dasgupta and Joseph E Stiglitz

34 Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition
Sherwin Rosen

56 A Diagrammatic Exposition of the Harberger Model with One Immobile Factor
Charles E McLure, Jr

83 An Economic Theory of Suicide
Daniel S Hamermesh and Neal M Soss

99 John Maynard Keynes Scientist or Politician?
Elizabeth Johnson

112 Group Segregation and Optimal Jurisdictions
Martin McGuire

133 Stability and the Quantity Theory
P W Howitt

Communications

152 The Interrelation between Labor and Capital Components of Racial Income Differences
George M von Furstenberg

163 An Optimal Program for Managing Common Property Resources with Congestion Externalities
Gardner Brown, Jr

174 Cognitive Range in the Theory of Revealed Preference
Anthony Y C Koo and Peter Schmidt

180 Consumers' Durable Goods, Human Capital, and the Pareto Optimality of the Lange-Lerner State
Michael R. Darby

185 A Test of the Permanent-Income Hypothesis
Comment by *James M Holmes*
Reply by *Prem S Laumas*

199 "Preferences, Separability, and the Patinkin Model" A Comment
Eitan Berglas and Assaf Razin

202 A Paradoxon of the Market Mechanism
Gerhard O Orosel

Miscellany

210 Drug-Law Enforcement Should We Arrest Pushers or Users?
Billy J Eatherly

215 **Puzzles and Problems**

217 **Book Reviews**

Robert A Mundell, *Monetary Theory Inflation, Interest, and Growth in the World Economy*
Don Patinkin

Trout Rader, *Theory of General Economic Equilibrium*
James Quirk

Nathan Rosenberg, *Technology and American Economic Growth*
Lars G Sandberg

Eugene Zaleski, *Planning for Economic Growth in the Soviet Union, 1918 1932*
R W Davies

Journal of Political Economy

Volume 82 Number 2, Part I March/April 1974

Articles

225 Private Savings Ultrarationality, Aggregation, and "Denison's Law"
Paul A David and John L Scadding

251 The Cost of School Time, Foregone Earnings, and Human Capital Formation
Donald O Parsons

267 Advertising and Profitability A Reappraisal
Harry Bloch

[illegible]nt Incentives and the Corporate Income Tax
[illegible]no

[illegible], Yield Curves, and Wage Settlements An Empirical [illegible]
[illegible] Comay, Arie Melnik, and Abraham Subotnik

[illegible]pretations of Departures from the Pareto Curve Firm-Size [illegible]ributions
[illegible] Ijiri and Herbert A Simon

333 Soviet Growth in Absence of Centralized Planning A Hypothetical Alternative
Micha Gisser and Paul Jonas

353 Ghetto Employment and the Model Cities Program
Bennett Harrison

373 Optimum City Size The External Diseconomy Question
J V Henderson

Communications

389 A Conditional Logit Model of Occupational Choice
Michael J Boskin

399 Occupational Licensing and the Public Interest
Alex Maurizi

414 On Equalizing the Distribution of Political Income
Dennis C Mueller, Robert D Tollison, and Thomas D Willett

423 Some Principles of Inflationary Finance
Donald A Nichols

431 A Bias in Current Measures of Economic Growth
Milton Friedman

Miscellany

433 Human Migration A Survey of Preclassical Literature
Robert Weinstein

437 **Book Reviews**

David M Gordon, *Theories of Poverty and Unemployment*
Sherwin Rosen

Leonard J Kirsch, *Soviet Wages Changes in Structure and Administration since 1956*
Murray Yanowitch

Michael J Webber, *Impact of Uncertainty on Location*
Jacques Thisse

Journal of Political Economy

Volume 82 Number 3 May/June 1974

Articles

443 International Capital Flows and Portfolio Equilibrium
Pentti J K Kouri and Michael G Porter

469 Taxation and the Rates of Return from Some U S Asset Holdings Abroad, 1960-1969
Herbert G Grubel

489 Cattle as Capital Goods and Ranchers as Portfolio Managers An Application to the Argentine Cattle Sector
Lovell S Jarvis

521 Rational Expectations and Bayesian Analysis
Richard M Cyert and Morris H DeGroot

537 Competitive Firm and the Theory of Input Demand under Price Uncertainty
Raveendra N Batra and Aman Ullah

549 Lifetime Earnings in a Professional Labor Market Academic Economists
George E Johnson and Frank P Stafford

Frank Knight in Retrospect

571 Frank Knight and the Tradition of Liberalism
Scott Gordon

578 Frank H Knight as Economic Theorist
Arthur H Leigh

Communications

587 Rational Response to the Money Supply
Richard Roll

598 The Honest Government's Guide to the Revenue from the Creation of Money
Leonardo Auernheimer

607 Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utilities A Note
Antonio Camacho and Jon Sonstelie

612 On Killing off the Market for Used Textbooks and the Relationship between Markets for New and Secondhand Goods
H Laurence Miller, Jr

620 A Note on Income (Labor) Inequality Income Tax Systems and Human Capital Theory
Claude Montmarquette

626 A Note on Equilibrium in Monopolistic Competition
Yoram Peles

631 Optimal Replacement of Capital Goods in Early New England and British Textile Firms A Comment
David Denslow, Jr , and David Schulze

638 Optimal Replacement of Capital Goods in Early New England and British Textile Firms Reply
Jeffrey G Williamson

641 A Note on the Stability of Metzler's Phase Diagram
Don Patinkin

645 Managerial Tenure under Private and Government Ownership in the Electric Power Industry
Louis De Alessi

654 Utility Maximization Sufficient for Competitive Survival Comment
Robert Piron

657 On the Economics of a University Toward the Determination of Marginal Cost of Teaching Services
Shlomo Brovender

Book Review

665 Philip Cagan, *The Channels of Monetary Effects on Interest Rates*
William Poole

Announcement from the Editors

In order to expedite the flow of manuscripts and to improve the quality of referee reports, the *JPE* is planning to initiate a policy of paying referees (in monetary form) The financing of these payments will require an increase in the submission fee, effective September 1, 1974, to $25 00 However, the fee will be refunded to authors whose papers are accepted for publication in the *Journal*

Journal of Political Economy

Volume 82 Number 4 July/August 1974

Articles

669 An Econometric Evaluation of Sulfur Taxes
James M. Griffin

689 Searching for the Lowest Price When the Distribution of Prices Is Unknown
Michael Rothschild

713 Karl Marx on the Economic Role of Science
Nathan Rosenberg

729 Advertising as Information
Phillip Nelson

755 Taxation and National Defense
Earl A Thompson

783 Vertical Control with Variable Proportions
Frederick W Warren-Boulton

Communications

803 The Risk Element in Occupational and Educational Choices Comment
John C Hause

809 Purchasing-Power Parity Theory and Economic Reform in Yugoslavia
Ryan C Amacher and John S Hodgson

817 Authoritarian versus Nonauthoritarian Approaches to Economic Development
G William Dick

829 Factor Proportions and Comparative Advantage in the Long Run Comment
Alan V Deardorff

835 The Value of Changes in the Probability of Death or Injury
Michael Jones-Lee

851 A Reexamination of the Kinky Oligopoly Demand Curve
Walter J Primeaux, Jr , and Mark R Bomball

863 Voucher Plans, Voting Models, and the Efficiency of Local School Finance
M L Eysenbach

873 Employer Costs and Discrimination The Case of Baseball
James Gwartney and Charles Haworth

883 Friedman versus Tobin A Comment
T Norman Van Cott and Gary Santoni

Miscellany

887 The Economics of Brushing Teeth
Alan S Blinder

892 Proxies and Dummies
Fritz Machlup

Review Article

893 The Cambridge-Cambridge Controversy in the Theory of Capital, A View from New Haven A Review Article
Joseph E Stiglitz

Announcement from the Editors

In order to expedite the flow of manuscripts and to improve the quality of referee reports, *JPE* is planning to initiate a policy of paying referees (in monetary form) The financing of these payments will require an increase in the submission fee, effective September 1, 1974, to $25 00 However, the fee will be refunded to authors whose papers are accepted for publication in the *Journal*

Benefit-Cost Analysis and Trade Policies

Partha Dasgupta
London School of Economics and Trinity Hall, Cambridge

Joseph E. Stiglitz
Cowles Foundation, Yale University

This paper extends the theory of optimal taxation and government production to open economies Appropriate rules for project evaluation and the determination of consumption, production, and trade taxes under a variety of restrictions (e g , less than 100 percent profit taxes, government budget constraint, foreign exchange constraint) are derived Among the results are (*a*) international prices should be used for evaluating public projects, unless there is a government budgetary constraint or there is a quota (this result does not require that tariff rates be optimally chosen), (*b*) no tariff should be levied on intermediates and only consumption taxes should be employed if there are 100 percent profit taxes If profits are not taxed at 100 percent, both consumption and trade taxes should be employed

1. Introduction

1 1 *Motivation*

In this paper we address ourselves to two questions· (1) In an open economy, in which the government is pursuing various trade and tax policies, and where some of the policy instruments are optimally chosen but the others are not, and in which, as a consequence, domestic price ratios differ from international price ratios, what is the appropriate relationship between international prices, domestic prices, and shadow

---

Research toward this paper was conducted while Dasgupta was at the Delhi School of Economics, Delhi, and Stiglitz was at the Institute of Development Studies, Nairobi We have gained much from conversations with James Mirrlees and Amartya Sen. Financial support from the Rockefeller Foundation, the National Science Foundation, and the Ford Foundation for Stiglitz's research is gratefully acknowledged.

prices in benefit-cost analysis? (2) Under what conditions is it optimal to impose taxes on imports and exports which are different from those imposed on domestically produced commodities, and what is the optimal tariff structure in these circumstances?

The motivation behind these questions is probably transparent Much of the recent discussion on benefit-cost analysis (in particular, Little and Mirrlees [1969] and Dasgupta, Marglin, and Sen [1972]) has been aimed at a detailed specification of the procedure that governments ought to follow in choosing industrial projects in the public sector But neither of the two documents referred to above presents a fully articulated model for arriving at the *detailed* recommendations It is, therefore, of some importance to check the kinds of circumstances under which these recommendations are indeed appropriate

The second set of questions that we are concerned with here, on the other hand, is closely related to the recent literature on effective protective rates (see, e g , Corden 1966, Ruffin 1969) The interest in this concept was presumably motivated by asking the question, What structure of tariffs leaves undistorted the *relative* prices of different commodities and factors of production? But the converse of the well-known proposition that in general equilibrium it is only relative prices that are determined is that, without changing relative prices, a (nonlump-sum) tax system can raise no revenue A more limited objective seems to have been the search for a tariff system which will leave undistorted relative commodity prices (but which will in general change relative factor prices and prices of commodities relative to factors) But it is pertinent to ask why one should be interested in a tariff structure that leaves relative commodity prices, or even relative supplies of different commodities, unchanged We know that most taxation is distortionary The relevant question is, If you must have distortions, what is the best set of distortions to have?[1]

1 2 The Open Economy

Some years ago Pigou (1947) attempted to extend to economies that trade the basic principles of optimal taxation that Ramsey (1927) had earlier described for economies that do not trade Ramsey had, for instance, argued that, in the case of independent demand and supply curves,[2] for small government revenue, taxes ought to be proportional to the sum of the inverse of the demand and supply elasticities.[3] Since the supply elasticity of foreign commodities is presumed in general to be greater than that for domestically produced commodities, Ramsey's analysis implied that, rather than impose a *surtax*, that is, impose a higher tax, on imports than on domestically produced goods, one should tax imports at a lower rate. But Pigou did not answer the question as to what ought to be done for commodities which are both domestically produced and imported. One would wish to know if the two activities ought to receive differential treatment.

---

[1] These are questions in the theory of the "second best," a subject of much discussion in recent years Although the first extensive discussion of "second-best" welfare economics was that of Meade (1955), the first real exercise in second-best welfare economics was probably that of Ramsey (1927) This paper is an extension to an open economy of the problems of optimal taxation discussed by Boiteux (1956), Diamond and Mirrlees (1971), Stiglitz and Dasgupta (1971), Dasgupta and Stiglitz (1972), and Mirrlees (1972)

[2] That is, where the demand and the supply for a commodity depends only on the price of that commodity (relative to, say, the price of labor)

There are, moreover, some important differences between taxation policy in closed economies and that in open economies, and it is not immediately apparent, when proper account of this is taken, whether Pigou's conclusions remain valid. Three differences come immediately to mind.

1. For imported goods, the producer surplus which accrues to producers of commodities which are not in perfectly elastic supply accrues to nationals of other countries and, accordingly, does not affect domestic welfare. For domestically produced goods it does.

2. Many countries face what is sometimes referred to as a "foreign exchange constraint." This implies that the earnings from export do not measure their social value, and the cost of imports does not measure the social cost. To what extent this should affect the rates at which different imports should be taxed, and whether this should result in governments' using, in benefit-cost analysis, prices that are different from those that rule at the border are questions that are still often debated. To see what is involved, consider an open economy in which there are no barriers to trade or to the flow of capital. If p_{it} is the international price of commodity i at time t, and z_{it} is the net import of commodity i at t, then trade balance requires only that

$$\sum_t \sum_i p_{it} z_{it} = 0 \tag{1.1}$$

A "foreign exchange" constraint imposes an additional condition: in any particular year, t, for instance, the deficit may not be larger than so much, say, ε_t. That is,[4]

$$\sum_i p_{it} z_{it} \leq \varepsilon_t \tag{1.2}$$

---

[3] Some of the assumptions implicit in Ramsey's analysis, e.g., the nontaxation of profits, have recently been made explicit (see Stiglitz and Dasgupta 1971).

[4] Perhaps a more realistic way of writing the constraint is that the cumulative deficit at time t be less than some number, $\hat{\varepsilon}_t$:

$$\sum_{\tau \leq t} \sum_i p_{i\tau} z_{i\tau} \leq \hat{\varepsilon}_t$$

This complicates the mathematics but does not affect the qualitative results given in section 7.

This is equivalent to saying that the country cannot borrow more than *s*, from abroad in any given year The foreign exchange constraint (*1.2*) is binding if and only if the country would like to borrow more than it can, which in turn implies that there is a constraint on the supply of investible funds (in other words, a savings constraint) If, as the analysis seems to suggest, the savings and foreign exchange constraints are in fact closely related, it is pertinent to ask whether it is fair to argue that the foreign exchange constraint does not affect relative prices used in benefit-cost analysis while the savings constraint does [5]

3 Although most governments do use excise and specific commodity taxes on domestic consumption to a limited extent, they are usually justified in terms of distributional objectives (e g , taxing luxuries, like perfumes), in terms of merit wants (e g , cigarettes and tobacco) or externalities, or as benefit taxes (e g , gasoline) On the other hand, almost all countries impose extensive import duties at varying rates, and both the commodities taxed at the border and the rates at which they are levied have probably less to do with welfare considerations than with the strength of various pressure groups

To put the matter differently, the conventional planning literature treats the "government" as if it were a single unit, the central planning agency being the basic coordinating body It simultaneously decides on taxes, tariffs, and investment projects Although this assumption has its merits—it allows one to pose a number of questions in their most pristine form—it is doubtful whether it makes for a good theory of government action [6] To take an example, suppose that project *X* requires equipment *E* as an input which it is optimal to import (since domestic production of *E* is too costly) If *E* is in fact imported, then assume *X* is a desirable project But the government project evaluator may know that if *X* is undertaken there will be powerful pressure groups at work forcing the departments responsible for levying tariffs to raise the tariff on *E* so as to enable private domestic producers to produce more of it domestically to meet the augmented demand for it What, then, should the decision be with regard to *X*? It does not seem apparent to us, at least, whether, in a closely interdependent economy, a nonoptimal tariff on one commodity should affect the tariffs and shadow prices of other commodities

Nevertheless, it has been argued that tariffs, whether optimally chosen or not, should generally be neglected for the purpose of benefit-cost analysis [7] One of the purposes of this paper is to investigate the kinds of circumstances under which this argument is correct.

---

[5] See Little and Mirrlees (1969) for a discussion of this

[6] See Sen (1972) for a lucid discussion of this range of issues

[7] See, in particular, Little and Mirrlees (1969)

In order to bring out the principles involved as clearly as possible, we consider in the following sections a sequence of models, in each of which the powers of the planner are somewhat more circumscribed than in the previous In section 2, the only constraint on the government is the impossibility of lump-sum taxation The government can impose 100 percent profit taxes and taxes on all production, consumption, and trade In section 3, we then consider the consequences of the government not being able to impose 100 percent profit taxes In section 4, we further restrict the government in allowing it to impose only trade taxes In section 5, some of the trade taxes are assumed given, and quotas are assumed to be imposed on some commodities, but the government must still make decisions concerning the choice of investment projects and the rates of tariffs on the remaining commodities In section 6, the consequences of trade policy partially dependent on project selection are explored In section 7, the consequences of a foreign exchange constraint are considered, while in section 8 those of a budgetary constraint on the government are investigated In section 9, we indicate briefly how distributional considerations may be integrated with our analysis Finally, in section 10, we summarize the major conclusions of the paper in terms of a set of rules [8]

Before setting out, we note three assumptions which will limit the generality of our conclusions (1) We shall assume that every commodity is either tradeable at a fixed international price or not tradeable at all. This not only simplifies the mathematics, but also allows us to separate the tax revenue-welfare effects of concern here with the terms of trade effects, which have been the focus of the optimal tariff literature so far (2) We shall assume that the only "distortions" in the economy are those introduced by government tax policy, otherwise, prices of factors and goods are competitively determined Accordingly, there is, for instance, no unemployment The presence of other distortions, for example, unemployment, undoubtedly would introduce discrepancies between market prices and shadow prices, whether there were tariff distortions or not [9] (3) For most of the analysis, we shall assume all individuals to be identical This means that we will not investigate the implications for tariffs and shadow prices of the interaction between savings and income distribution

---

[8] The two authors who have come closest to discussing the issues on tariff policy presented here are Meade (1955) and Ramaswami and Srinivasan (1968) The model considered by Ramaswami and Srinivasan may be viewed as a special case of that presented here, for they are concerned exclusively with an economy in which only border taxes can be levied (a circumstance we discuss in section 4) Moreover they do not allow for the possibility of public production Their conclusion that duties need not be uniform in a second-best revenue tariff is reestablished here (section 4 1) and the precise formulae for the optimum tariffs are given For an excellent survey see Bhagwati (1971)

[9] For an analysis of this class of situations see Marglin (1972) and Stiglitz (1973, in press)

All of these raise important problems which we hope to pursue elsewhere

2. The Fully Controlled Economy

In this section, we consider an economy in which there are no foreign exchange constraints and no trade and domestic distortions apart from those which the government *deliberately* imposes The only constraint imposed upon the government is the unfeasibility of its employing lump-sum taxes The government's problem is to decide what taxes to impose on domestic producers, what tariffs to levy on imports (both intermediate and final commodities), what excise taxes to set on domestic consumption, and which projects to undertake It wishes to do this in such a way that all markets clear, trade balances, and welfare is maximized

2 1 *Notation*

We introduce the following notation

c_i = consumption of the ith commodity (for factors supplied, $c_i < 0$),[10]
y_i = production of commodity i by domestic private industry (for factors used by private industry $y_i < 0$),
y_i^j = production of commodity i by the jth private producer,
q_i = consumption price of commodity i,
p_i = producer price of commodity i When different producers face different prices for the same commodity, the price is denoted as p_i^j,
z_i = net import of commodity i If it is a net export, then $z_i < 0$,
x_i = net output of commodity i in the public sector If it is a net input, then $x_i < 0$,
t_i = tax on the import of commodity i,
Γ_i = tax on producers of commodity i (when different producers face different taxes, the tax is denoted by Γ_i^j),
τ_i = difference between the international price for commodity i and the domestic producer price

Where there is no risk of ambiguity we shall denote a vector, say (c_i), simply by c The scalar product of two vectors, say q and c, will be denoted by $q \cdot c$ Occasionally we shall denote by $[\eta]$ a square matrix with typical element η_{ik}

---

[10] These commodities may be time dated, so the results are valid for the intertemporal case as well as the static case

Not all commodities are tradeable, neither are all commodities consumable. We shall denote by A the set of all commodities in the economy; by T the set of all tradeable commodities; by N the set of all nontradeable commodities; by D the set of all domestically produced commodities (some of which still might be traded); by I the set of all intermediate goods, and by C the set of all consumption goods. Thus $A = C \cup I = T \cup N$, $C \cap I = \emptyset$, $T \cap N = \emptyset$, $N \subset D$. Our numeraire good will be indexed as $i = 0$. For notational simplicity, we shall suppose, unless noted otherwise, that it is at the same time traded, domestically produced, and consumed Our units are chosen so that the international price of every traded commodity is unity

2 2 *Basic Relations*

There are a few simple relations among the variables defined in section 2 1 which we set forth here To begin with, the commodity balance equations are

$$c_i = y_i + x_i + z_i, \tag{2 1}$$

where it is understood that $z_i = 0$ for $i \in N$, $c_i = 0$ for $i \in I$, and $x_i = y_i = 0$ for $i \notin D$

The total output y_i from domestic private firms is

$$y_i = \sum_j y_i^j \qquad i \in D \tag{2 2}$$

Now domestic consumption prices equal international prices plus the taxes on traded goods We then have

$$q_i = p_i + \Gamma_i = 1 + t_i \qquad i \in C \cap T \cap D, \tag{2 3a}$$

$$q_i = p_i + \Gamma_i \qquad i \in C \cap N, \tag{2 3b}$$

$$q_i = 1 + t_i \qquad i \in C \cap T \text{ and } i \notin D, \tag{2 3c}$$

$$p_i + \Gamma_i = 1 + t_i \qquad i \in T \cap D \cap I \tag{2 3d}$$

In other words,

$$\Gamma_i - t_i = \tau_i = 1 - p_i \qquad i \in T \cap D \tag{2 3e}$$

Thus, $t_i > 0$ for a tariff on an import or a subsidy on an export, $t_i < 0$ for a tariff on an export (i e , an export tax) or a subsidy on an import, $\tau_i > 0$ for a production tax in excess of the tariff, $-t_i < \tau_i < 0$ for a production tax that is less than the tariff If $\tau_i = -t_i$, domestic production is neither taxed nor subsidized ($q_i = p_i$)

2.3 *Consumer Behavior*

Throughout the discussion (except in section 9) we shall suppose that all individuals are identical. We shall, therefore, talk of the "representative" individual. He is assumed to maximize his utility,

$$\max U(c), \tag{2.4}$$

subject to his budget constraint,

$$q \cdot c = M, \tag{2.5}$$

where M is his income. If there are no lump-sum taxes or subsidies, then M is just equal to profits (pure rents) after tax. If π^j denotes the net profit of the jth firm, and τ_π denotes the tax rate on profits, then

$$M = \sum_j (1 - \tau_\pi)\pi^j = (1 - \tau_\pi)\pi, \text{ where } \pi = \sum_j \pi^j \tag{2.6}$$

The solution to (2.4) may be written as

$$V(q, M) = \max U, \tag{2.7}$$

where V is known as the consumer's indirect utility function.

2.4 *Behavior of Firm*

There are m different private firms in the economy. For simplicity, we assume that all firms face strictly concave production functions.[11]

$$F^j(y^j) = 0 \qquad j = 1, \ldots, m \tag{2.8}$$

The jth firm maximizes profits at the given price vector p^j.

$$\max p^j \cdot y^j \qquad \text{subject to } F^j(y^j) = 0 \tag{2.9}$$

The solution to (2.9) is denoted by $\pi^j(p^j)$.[12] Moreover, in equilibrium it will be the case that

$$p_i^j/p_k^j = \frac{\partial F^j/\partial y_i^j}{\partial F^j/\partial y_k^j} \qquad i, k \in D \tag{2.9a}$$

These equations can be solved for the supply of commodities (demands for factors) as a function of p^j.[13]

$$y_k^j = y_k^j(p^j) \tag{2.9b}$$

---

[11] As we have shown elsewhere (Dasgupta and Stiglitz 1972) there is no important loss of generality by assuming strict concavity. See Dasgupta and Stiglitz (1972) and Mirrlees (1972) for a discussion of some possible anomalies.

[12] It has become customary in the literature on general equilibrium to refer to π^j as the net profit for the jth firm (see, in particular Debreu [1959]), whereas, of course, it denotes what has traditionally been referred to as "pure rents" accrued by the firm. We stick to the modern usage here largely because it sounds more reasonable to say that a firm is concerned with maximizing profits rather than its rents.

[13] Note that, if we had a constant returns to scale function, then we could solve for relative factor intensities, but not for total factor demands without knowing quantities produced. See, however, Dasgupta and Stiglitz (1971).

Indeed, it is a well-known result that

$$\frac{\partial \pi^j}{\partial p_k^j} = y_k^j \tag{2 9c}$$

One other fact will be of use Differentiating the production constraint, we obtain

$$\sum \frac{\partial F^j}{\partial y_i^j} \frac{\partial y_i^j}{\partial p_k^j} = \frac{\partial F^j / \partial y_0^j}{p_0^j} \sum p_i^j \frac{\partial y_i^j}{\partial p_k^j} = 0. \tag{2 9d}$$

2 5 *Trade Balance*

The condition for the balance of trade may be written as

$$B = \sum_{k \in T} z_k = \sum_{k \in T} (c_k - y_k - x_k) = 0 \tag{2 10}$$

Thus, we can view B as a function of q, p, and x

$$\frac{\partial B}{\partial q_i} = \sum_{k \in T} \frac{\partial c_k}{\partial q_i}, \tag{2 10a}$$

$$\frac{\partial B}{\partial p_i} = y_i \sum_{k \in T} \frac{\partial c_k}{\partial \pi} - \sum_{k \in T} \frac{\partial y_k}{\partial p_i}, \tag{2 10b}$$

$$\frac{\partial B}{\partial x_k} = -1 \tag{2 10c}$$

2 6 *The Government*

The government's production function we shall denote by

$$G(x) = 0 \tag{2 11}$$

As equation (2 11) indicates, we are tacitly supposing that the public sector is concerned with the transformation of the same set of goods as the domestic private sector is We assume this simply for notational ease [14]

---

[14] We are supposing, then, that there are no public goods in this economy If we were to assume that the public sector produces public goods as well we would obtain one set of first-order conditions in addition to the sets of conditions obtained below This additional set would dictate the optimum supply of public goods For a discussion of the implications of such a condition see Diamond and Mirrlees (1971) and Stiglitz and Dasgupta (1971). Our concern here is with obtaining shadow prices of private goods, and it is trivial to show that the results derived here apply without modification to the case where public goods are also produced by the government For this reason we ignore public goods here Similarly, there may be several production units in the public sector But the result derived earlier in Boiteux (1956), Diamond and Mirrlees (1971), and Stiglitz and Dasgupta (1971), that the public sector should always be productively efficient, remains valid. To keep the notation simple we have therefore amalgamated all public sector production units into one single unit

We shall suppose that the government controls trade and private production, but only indirectly through tax and tariff policies. The problem before the planning agency is straightforward: it wishes to maximize the welfare V of the representative consumer subject to the private sector production possibility curves (2 8), the public sector production possibility curve (2 11), the balance of payments condition (2 10), and the market clearing equations for the nontraded goods The government's controls are direct government production, tariffs, and consumption, production, and profits taxes But it is easy to establish (using [2 3]) that controlling taxes and tariffs is essentially equivalent to controlling producer and consumer prices, and it turns out that the latter are conceptually and analytically easier to use Similarly, although controlling p, q, and τ determines z, so that the only constraint on the traded goods is the balance of payments constraint (i e , we do not need to impose separate market clearing constraints on each of the traded commodities), it is convenient to formulate the problem as if z were control variables and impose the additional market clearing equation for each of the traded goods[15] Finally, we note (see Stiglitz and Dasgupta 1971) that optimality requires the government to set $\tau_\pi = 1$, provided the government requires some resources to be raised by distortionary taxes In view of the fact that the analysis is trivial in those cases where the government does not need to impose distortionary taxes to raise its revenue, we set $\tau_\pi = 1$ whenever it is assumed to be feasible For the remainder of this section we suppose that it is

Finally, we normalize by setting $q_0 = 1$ and $p_0^j = 1$ for all j[16]

---

[15] In other words, the problem

$$\max_{\{x\ p,q\}} V[q, \pi(1 - \tau_\pi)] + \lambda\left[\sum_{k\in T}(x_k + y_k - c_k)\right] + \sum_{k\in N}\rho_k(x_k + y_k - c_k) + \mu G + \sum_{j=1}^{m}\psi_j F^j(y^j)$$

and the problem

$$\max_{\{x,p,q,z\}} V[q, \pi(1 - \tau_\pi)] - \lambda\sum z_k + \sum_{k\in A}\rho_k(z_k + x_k + y_k - c_k) + \mu G + \sum_{j=1}^{m}\psi_j F^j(y^j)$$

are equivalent When the outputs and inputs are viewed as functions of producer prices, the production constraints (2 8) are already embedded in the problem (through 2 9), and hence we could omit the term

$$\sum_{j=1}^{m}\Psi_j F^j(y^j)$$

in our Lagrangian expression In the centralized problem, where outputs and inputs are considered as direct controls, this term must be retained For purposes of symmetry with the centralized problem, we retain the production constraints explicitly within our Lagrangian expression

[16] For a discussion of these normalizations, see Dasgupta and Stiglitz (1972)

2 7. *Optimal Public Policy*

We now form the Lagrangian of the problem·[17]

$$\mathscr{L} = V(q, 0) - \lambda B + \mu G + \sum_{j=1}^{m} \psi_j F^j(y^j) + \sum_{k \in A} \rho_k \left(\sum_{j=1}^{m} y_k^j + x_k + z_k - c_k \right) \qquad (2\ 12)$$

We obtain the first-order conditions

$$\frac{\partial V}{\partial q_i} - \sum_{k \in C} \rho_k \frac{\partial c_k}{\partial q_i} = 0 \qquad i \in C \text{ and } i \neq 0 \qquad (2\ 13)$$

$$\sum_{k \in D} \left(\psi_j \frac{\partial F^j}{\partial y_k^j} + \rho_k \right) \frac{\partial y_k^j}{\partial p_i^j} = 0 \qquad i \in D \text{ and } i \neq 0, \quad j = 1, \quad , m, \qquad (2\ 14)$$

$$\mu \frac{\partial G}{\partial x_i} + \rho_i = 0 \qquad i \in D, \qquad (2\ 15)$$

and

$$-\lambda + \rho_i = 0 \qquad i \in T \qquad (2\ 16)$$

From equations (2 14–2 16) we obtain the now-familiar result [18]

$$\frac{\partial G/\partial x_i}{\partial G/\partial x_0} = \frac{\rho_i}{\rho_0} = \frac{\partial F^j/\partial y_i^j}{\partial F^j/\partial y_0^j} = p_i^j = 1 \qquad i \in T \cap D, \quad j = 1, \quad , m, \qquad (2\ 17\text{a})$$

and

$$\frac{\partial G\ \partial x_i}{\partial G/\partial x_0} = \frac{\partial F^j{}_{,}\partial y_i^j}{\partial F^j/\partial y_0^j} = \frac{\rho_i}{\lambda} = p_i \qquad i \in N \subset D \qquad (2\ 17\text{b})$$

Equations (2 17a) and (2 17b) describe the desirability of overall production efficiency for the economy All production units in the economy ought to use international prices in selecting their optimum techniques, and there ought to be no taxes or tariffs on intermediate goods [19]

It is then apparent that, at least in this case, Pigou's (1947) conclusion, that domestic production ought to be taxed at a higher rate than imports, is incorrect No differentiation between domestic and foreign production should be made

---

[17] We recall the convention adopted in discussing (2 1) that it is understood that $z_i = 0$ for $i \in N$, $c_i = 0$ for $i \in I$, $y_i = 0$, and $x_i = 0$ for $i \notin D$

[18] The second equality follows from observing that, for given j, we have, with (2 9d), as many equations of the form (2 14) as commodities, since we thus have a system of n homogenous equations in n unknowns, we require $\psi_j(\partial F^j/\partial y_k) + \rho_k = 0$ See Dasgupta and Stiglitz (1971) for a more detailed discussion

[19] This is the central result in Diamond and Mirrlees (1971) For a thorough exploration of the application of this result see Little and Mirrlees (1969)

Turning to equation (2.13), it is now simple to show that it reduces to[20]

$$\sum_{k\in T\cap C}\frac{t_k}{c_i}\left(\frac{\partial c_i}{\partial q_k}\right)_{\bar{U}}+\sum_{k\in N\cap C}\frac{\Gamma_k}{c_i}\left(\frac{\partial c_i}{\partial q_k}\right)_{\bar{U}}$$

$$\equiv-\sum_{k\in T\cap C}\frac{t_k}{q_k}\eta^d_{ik}-\sum_{k\in N\cap C}\frac{\Gamma_k}{q_k}\eta^d_{ik}=-\theta\qquad i\in C,\qquad(2.18)$$

where

$$\theta\equiv1-\frac{1}{\lambda}\frac{\partial V}{\partial M}-\sum_{k\in T\cap C}t_k\frac{\partial c_k}{\partial M}-\sum_{k\in N\cap C}\Gamma_k\frac{\partial c_k}{\partial M},$$

and

$$\eta^d_{ij}\equiv-\left(\frac{\partial\ln c_i}{\partial\ln q_j}\right)_{\bar{U}}$$

It follows from equation (2.18) that for small taxes the consumption of all commodities is reduced by the same amount (from what it would have been had producer prices been charged).[21]

3. Limited Profit Tax on Private Firms

3.1 General Considerations

For a number of reasons, no government imposes a 100 percent tax on profits. This requires a good deal of modification in the analysis. Now profits enter directly into the utility function, and changes in profits affect the demands for various goods. For simplicity, we consider the

---

[20] To confirm the validity of equation (2.18) one uses (2.3a), (2.3b), (2.16), and (2.17b) in equation (2.13), and noting that $\partial V/\partial q_i=-c_i(\partial V/\partial M)$, one obtains

$$\frac{-c_i}{\lambda}\frac{\partial V}{\partial M}-\sum_{k\in C\cap T}(q_k-t_k)\frac{\partial c_k}{\partial q_i}-\sum_{k\in C\cap N}(q_k-\Gamma_k)\frac{\partial c_k}{\partial q_i}=0\qquad i\in C$$

On using Slutsky's equation, the fact that the Slutsky matrix is symmetric, *and* the fact that

$$\sum_{k\in C}q_k\left(\frac{\partial c_k}{\partial q_i}\right)_{\bar{U}}=0,$$

the foregoing equation reduces, on collecting terms, to the form

$$\sum_{k\in C\cap T}\frac{t_k}{c_i}\left(\frac{\partial c_i}{\partial q_k}\right)_{\bar{U}}+\sum_{k\in C\cap N}\frac{\Gamma_k}{c_i}\left(\frac{\partial c_i}{\partial q_k}\right)_{\bar{U}}$$

$$=\frac{\partial V/\partial M}{\lambda}-\sum_{k\in C}q_k\frac{\partial c_k}{\partial M}+\sum_{k\in C\cap T}t_k\frac{\partial c_k}{\partial M}+\sum_{k\in C\cap N}\Gamma_k\frac{\partial c_k}{\partial M}$$

which immediately yields equation (2.18).

[21] Note the difference between this result and the underlying notions of the effective tariff literature. There, the focus is on a tariff structure which reduces production of each commodity by an equal percentage from the pretax situation and the analysis often ignores the change in factor prices of nontraded inputs; here, we reduce consumption by an equal percentage from what it would have been with the new producer prices; for traded goods, the only change in production results from the change in prices of nontraded inputs.

polar case where no tax on profits can be imposed. The intermediate cases are trivial to work out.

The Lagrangian now takes the form

$$V\left[q, \sum_{j=1}^{m} \pi^j(p^j)\right] - \lambda \sum_{k \in T} z_k + \mu G(x) + \sum_{j=1}^{m} \psi_j F^j(y^j)$$
$$+ \sum_{k \in A} \rho_k \left(\sum_{j=1}^{m} y_k^j + x_k + z_k - c_k\right) \qquad (3.1)$$

The first-order conditions now read as

$$\frac{\partial V}{\partial q_i} - \sum_{k \in C} \rho_k \frac{\partial c_k}{\partial q_i} = 0 \qquad i \in C, \quad i \neq 0, \qquad (3.2)$$

$$\frac{\partial V}{\partial M}\frac{\partial \pi^j}{\partial p_i^j} + \sum_{k \in D}\left(\psi_j \frac{\partial F^j}{\partial y_k^j} + \rho_k\right)\frac{\partial y_k^j}{\partial p_i^j} - \sum_{k \in C} \rho_k \frac{\partial c_k}{\partial M}\frac{\partial \pi^j}{\partial p_i^j} = 0$$
$$j = 1, \ldots, m, \quad i \in D \text{ and } i \neq 0, \qquad (3.3)$$

$$\mu \frac{\partial G}{\partial x_i} + \rho_i = 0 \qquad i \in D, \qquad (3.4)$$

$$-\lambda + \rho_i = 0 \qquad i \in T \qquad (3.5)$$

3.2 *All Commodities Tradeable*

Consider, for simplicity, the case where all goods are tradeable. Using equation (3.5), we can write equation (3.2) as

$$\frac{\partial V}{\partial q_i} - \lambda \sum_{k \in C} \frac{\partial c_k}{\partial q_i} = 0 \qquad i \in C \text{ and } i \neq 0 \qquad (3.2a)$$

Equation (3.2a) has an immediate interpretation. By raising tariffs (consumer prices), we improve the balance of payments, by discouraging imports. The "cost" of this is, of course, a decrease in utility—higher prices for consumer goods makes one worse off. There is, therefore, a trade-off between "trade surplus" and welfare. Equation (3.2a) reflects the fact that the marginal rate of substitution between the two should be the same regardless of which price we vary, that is, the loss in welfare per unit gain in the balance of payments surplus should at the margin be the same for all commodity prices.[22] Turning to equation (3.3), by using (2.9c) and the fact that

$$\sum_{k \in D} \frac{\partial F^j}{\partial y_k^j}\frac{\partial y_k^j}{\partial p_i^j} \equiv 0 \qquad \text{all } i,$$

---

[22] Exactly the same reasoning applies to exports. An export tax lowers the domestic price, increases welfare, and, because it increases consumption of the given commodity, makes the balance of payments situation worse.

we obtain

$$y_i^j \frac{\partial V}{\partial M} = \lambda \left(y_i^j \sum_{k \in C} \frac{\partial c_k}{\partial M} - \sum_{k \in D} \frac{\partial y_k^j}{\partial p_i^j} \right) \qquad (3.3a)$$

Equation (3.3a) has an interpretation similar to (3.2a). An increase in the producer price of an output increases profits and makes the consumer better off. It has two effects on the balance of payments: by increasing the supply of the commodity it is improved, but by increasing the demand (as a result of the higher income from the higher profits) it is deteriorated. Equation (3.3a) reflects the fact that at the optimal point the latter effect on the balance of payments must dominate the former, and indeed, the marginal rate of substitution of gains in welfare and the worsening of the balance of payments must be the same, regardless of which producer price is varied.

Equations (3.4) and (3.5) have the following similar interpretation: the marginal rate of transformation between two commodities in public production must be equal to their relative marginal effects on the balance of payments at fixed producer and consumer prices. Thus, $\partial B/\partial x_i = 1$, the government project evaluator should use international prices in selecting public sector projects.

To see more precisely what equations (3.2a) and (3.3a) imply for the structure of tariffs and taxes, we begin by noting that equation (3.2a) reduces to the form

$$\sum_{k \in C} \frac{t_k}{c_i} \left(\frac{\partial c_i}{\partial q_k} \right)_{\bar{U}} = -\theta < 0 \qquad \text{where } i \in C \qquad (3.6)$$

This is the familiar formula for the optimal tax structure obtained earlier as equation (2.18). It implies that taxes (tariffs) ought to be such that consumption of all commodities is reduced by the same percentage along the compensated demand curve from what it would have been had international prices been charged.

To see what equation (3.3a) implies, we note that, from it, we can derive the following relationship:[23]

$$\sum_{k \in D} \frac{\tau_k^j}{y_i^j} \frac{\partial y_i^j}{\partial p_k^j} = -\sum_{k \in C} \frac{t_k}{c_\ell} \left(\frac{\partial c_\ell}{\partial q_k} \right)_{\bar{U}} = \theta > 0,$$

$$j = 1, \ldots, m, \quad i \in D \text{ and } \ell \in C \qquad (3.7)$$

---

[23] Substituting (3.2a) into (3.3a), and using the fact that $\partial V/\partial q_\ell = -c_\ell(\partial V/\partial M)$, we obtain

$$-\frac{1}{c_\ell} \sum_{k \in C} \frac{\partial c_k}{\partial q_\ell} = \sum \frac{\partial c_k}{\partial M} - \frac{1}{y_i^j} \sum \frac{\partial y_k^j}{\partial p_i^j}$$

Using the Slutsky equation, we obtain

$$\frac{1}{c_\ell} \sum_{k \in C} \left(\frac{\partial c_k}{\partial q_\ell} \right)_{\bar{U}} = \sum_{k \in D} \frac{\partial y_k^j}{\partial p_i^j} \frac{1}{y_i^j}$$

Equation (3 7) implies that the percentage change in the jth firm's output of the ith commodity or input of the ith factor should be the same for all firms, commodities, and factors. Notice that we have included for the possibility of intermediate goods Therefore, as far as producer taxation is concerned, one observes that intermediate goods are treated in no way different from final consumer goods[24]

To see more clearly what equation (3 7) implies for the structure of producer taxes, consider the case where we treat all the private producers identically and, hence, can aggregate them together Define

$$\hat{t}_k \equiv \frac{\tau_k}{1 - \tau_k} = \frac{\tau_k}{p_k} \qquad \text{and } \eta^s_{ik} \equiv \frac{p_k}{y_i}\frac{\partial y_i}{\partial p_k} \tag{3 8}$$

Then equation (3 7) can be written in the matrix form

$$\hat{t} \cdot [\eta^s] = \theta E, \qquad \text{or } \hat{t} = \theta E \ \ [\eta^s]^{-1}, \tag{3 9}$$

where $E = (1, \quad , 1)$, a vector with unity everywhere

Equation (3 9) provides a simple expression for calculating the optimal structure of producer taxes Notice that only in the limiting case of infinite supply elasticities (constant returns to scale) is it optimal to impose no tax on domestic production

Since a tax-tariff structure affects only relative prices, whether there is a tax or subsidy on a particular commodity depends on what we choose as our numeraire, that is, the price of a given commodity will rise relative to that of some other commodities, fall relative to still others It follows that Pigou's (1947) conjecture that the tax on the output of a given commodity by a domestic producer ought to be higher than the tax on the import of the same commodity, is not, in fact, a well-formulated one[25] A natural choice of numeraire for a small country with a principal export crop is that export crop Then, if the final good is produced by the intermediate good alone, $\hat{t}_f = 0$ (by definition), $\hat{\tau}_i = -\theta/\eta^s_f(1 - \pi/p_f y_f) < 0$ (where η^s_f is the supply elasticity of the final good, p_f and y_f are its price and output), that is, a production subsidy (relative to the tariff on the good) is imposed on the intermediate good

---

Finally, since

$$\sum_{k \in D} p^j_k \frac{\partial y^j_k}{\partial p^j_i} = \sum_{k \in C} q_k \left(\frac{\partial c_k}{\partial q_i}\right)_{\bar{U}} = 0 \cdot \quad \text{when } \ell \in C \quad \text{and} \quad i \in D$$

and $\partial y^j_k/\partial p^j_i = \partial y^j_i/\partial p^j_k$, $i, k \in D$, $(\partial c_k/\partial q_\ell)_{\bar{U}} = (\partial c_\ell/\partial q_k)_{\bar{U}}$, $\ell, k \in C$, using (2 3), we obtain (3 7)

[24] Our general formulation allows us to tax imports of intermediate goods at a different rate from domestic production of intermediate goods, and this in fact will in general be desirable, since there will be different elasticities associated with each

[25] The question may be reformulated, as follows Is it true that $q_i/q_0 > 0$ implies $p_i/p_0 < 1$? Consider a change of numeraire to j Assume that $q_j/q_0 > 1$, $p_j/p_0 < 1$ Then the signs of $(q_i/q_j) - 1$ and $(p_i/p_j) - 1$ are ambiguous In particular, it is possible for $q_i/q_j > 1$ while $p_i/p_j > 1$

3.3 *Social Valuation of Private Projects*

In many developing countries, the government also authorizes the construction of projects in the private sector. Much of benefit-cost analysis is directed to an evaluation of these "private" projects. Assume that no restrictions are imposed on the government in its tax treatment (except the levying of profit taxes). Then, from (3.3a), we note that whether the kth small project increases welfare depends on the sign of

$$\frac{\partial V}{\partial M}\pi^k + \lambda\left(\sum \Delta y^k - \sum \Delta c^k\right), \tag{3.10}$$

where Δc^k is the induced increase in consumption. The first term is the profits, evaluated at domestic prices, the second term is the effect of the project on the balance of payments. Thus, profits at domestic prices must exceed the total foreign exchange cost of the project evaluated at the shadow price of foreign exchange, $\lambda/(\partial V/\partial M)$.

This may be reformulated since

$$\begin{aligned} \sum q_i \Delta c_i^k &= \pi^k, \\ \sum \Delta y^k - \sum \Delta c^k &= \sum \Delta y^k - \pi^k + \sum t_i \Delta c_i^k \\ &= \sum \tau_i^k \Delta y_i^k + \sum t_i \Delta c_i^k \equiv R^k, \end{aligned}$$

where R^k is the change in government revenue from trade taxes from the project. Thus, we require $\pi^k + [\lambda/(\partial V/\partial M)]R^k \geq 0$. Profits at domestic prices, plus changes in government trade tax revenues, evaluated at the shadow price of foreign exchange, $[\lambda/(\partial V/\partial M)] > 1$, must be nonnegative.[26] This makes clear that the criteria sometimes advocated, namely, $\pi^k + R^k \geq 0$, or indeed the criterion $\Sigma \Delta y^k > 0$ (that profits evaluated at international prices be nonnegative), are both overly restrictive. The point is that profits and government revenue should not be evaluated at the same set of shadow prices.

3.4 *Nontradeables*

When there are nontradeables, we must include in our calculations the change in consumption and production of these commodities. As equations (3.4) and (3.5) make clear, nontradeables are to be evaluated in terms of their marginal foreign exchange costs. That is,

$$\rho_i = \lambda \frac{\partial G/\partial x_i}{\partial G/\partial x_k} \qquad k \in T \text{ and } i \in N \tag{3.11}$$

---

[26] It may be noted that λ is the shadow price of foreign exchange in terms of utility numeraire, so that $\lambda/(\partial V/\partial M)$ is the shadow price of foreign exchange in terms of domestic income.

The shadow price of a nontradeable (in utility numeraire) is, therefore, equal to the value (at international prices) of the foreign exchange which would be generated were one unit less of it produced domestically, and instead more units of the tradeables were produced, valued at the shadow price of foreign exchange (in utility numeraire)

The rest of the analysis proceeds easily One obtains from equations (3 2) and (3 3) tax formulae which are similar to equations (3 6) and (3 7), except now, for the nontradeables

$$t_i = q_i - \frac{\rho_i}{\lambda} \qquad i \in N \cap C, \qquad (3\ 12a)$$

$$\tau_i = \frac{\rho_i}{\lambda} - p_i \qquad i \in N \qquad (3\ 12b)$$

which give the difference between the consumer price and the shadow price (eq [3 12a]) and the difference between the shadow price and the producer price (eq [3 12b]) If we now suppose that there is a single producer of a given nontraded good (or that the producers of the good can be aggregated) and also assume that the demand and supply curves are independent of other prices, we find that the tax levied on a unit of a nontraded good is

$$q_i - p_i = t_i + \tau_i = \theta\left[\frac{1}{\eta^d_{ii}} + \frac{1}{\eta^s_{ii}}\right] \qquad i \in N \cap C \qquad (3\ 13)$$

The formula given in (3 13) is a generalization of the result in Ramsey (1927) [27]

3 5 *Foreign Firms*

In the foregoing analysis we assumed that profits accrued to the nationals of the country in which production occurred In most developing countries, a significant part of the profits is repatriated to some foreign country Consider then the case where all profits of a given firm are repatriated These profits do not affect the welfare of the representative consumer and, indeed, contribute to the deterioration of the balance of payments situation In this situation, the balance of payments condition reads as

$$\sum_{k \in T} z_k + \sum_{j \in F} \pi^j = 0,$$

where F is the set of foreign firms Suppose again, for ease of exposition, that all goods are tradeable For $j \in F$ we note that the first-order condition

---

[27] It would perhaps have been more natural to choose as our numeraire labor, a nontraded "good" Then, assume we have two methods of obtaining a commodity, one by producing it with labor, the other by importing This is the case that Pigou (1947) probably had in mind, and a simple recasting of our results yields the conclusion that a surtax is imposed on the domestic production of the commodity.

(3 3) can be written as

$$-\lambda \frac{\partial \pi^j}{\partial p_i^j} + \sum_{k \in D} \left(\psi_j \frac{\partial F^j}{\partial y_k^j} + \rho_k \right) \frac{\partial y_k^j}{\partial p_i^j} = 0, \qquad j \in F \text{ and } i \in D, \quad i \neq 0 \tag{3.14}$$

On using equations (3 4), (3 5), and (2 9a), in equation (3 14) one obtains

$$-\lambda y_i^j + \psi_j \frac{\partial F^j}{\partial y_0^j} \sum_{k \in D} p_k^j \frac{\partial y_k^j}{\partial p_i^j} + \lambda \sum_{k \in D} \frac{\partial y_k^j}{\partial p_i^j} = 0,$$

or using (2 9d)

$$\sum_{k \in D} \frac{\partial y_k^j}{\partial p_i^j} - y_i^j = 0 \qquad i \in D \text{ and } j \in F$$

An increase in the producer price has, therefore, two effects it increases domestic production, thus reducing the need for imports and improving the balance of payments, and it increases profits, which worsens the balance of payments (when repatriated) Optimality requires at the margin that two effects offset each other

Again using (2 9d) and (2 3), we obtain

$$\sum_{k \in D} \frac{\tau_k^j}{y_i^j} \frac{\partial y_i^j}{\partial p_k^j} = 1 \qquad i \in D \text{ and } j \in F \tag{3 15}$$

Writing $\hat{\tau}_k^j = \tau_k^j / p_k^j$ and $\eta_{ik}^{sj} = (\partial y_i^j / \partial p_k^j)(p_k^j / y_i^j)$, one obtains from equation (3 15) the formula

$$\hat{\tau}^j = E \; [\eta^{sj}]^{-1} \tag{3 16}$$

One should observe that, as before, industries which have more inelastic supply curves should be taxed at a higher rate More interesting and more important is the fact that equation (3 16) implies that the tax rates are independent of the government's need for revenue (This is in contrast with the case of domestically owned firms, where the tax rates depend on θ)[28] That is to say, the tax rates on the commodities produced by foreign-owned firms are determined completely by balance of payments considerations This results in the tax rates depending only on the properties of the production function of the industry

The cost-benefit criterion for accepting projects of foreign-owned corporations must similarly be modified The criterion should be simply whether $\lambda(-\pi^k + \sum \Delta y^k) = \lambda[\sum(1 - p_i) \Delta y_i^k] = \lambda(\sum \tau_i \Delta y_i^k) > 0$, that is, whether the "surtax" (subsidy) on domestic production yields net tax revenue (This assumes that the project will not affect prices facing other firms)

---

[28] Compare equation (3 7)

4. Inability of Taxing Domestic Production

4.1 No Profit Taxation

We now modify the model somewhat to assume that the government cannot impose producer taxes. We suppose that the only taxes that are feasible are trade taxes. Domestic producers then face the same price vector as does the consumer. To keep the notation simple, we suppose initially that all goods are tradeable, consumable, and domestically produced. That is $A = D = T = C$. We can again formulate the problem as if the controls at the disposal of the government are the domestic price vector, q, the vector of public production, x, and the vector of trade flows, z.[29] (Note that we must have $q = p = 1 + t$.) Using the Lagrangian expression (3.1), we now obtain the first-order conditions as

$$\frac{dV}{dq_i} + \sum_{j=1}^{m} \sum_{k \in A} \left(\frac{\partial F^j}{\partial y_k^j} \psi_j + \rho_k \right) \frac{\partial y_k^j}{\partial q_i} - \sum_{k \in C} \rho_k \frac{dc_k}{dq_i} = 0 \qquad i \in A \text{ and } i \neq 0, \tag{4.1}$$

$$\mu \frac{\partial G}{\partial x_i} + \rho_i = 0 \qquad i \in A, \tag{4.2}$$

$$-\lambda + \rho_i = 0 \qquad i \in A, \tag{4.3}$$

where dV/dq_i and dc_k/dq_i are to be regarded as total derivatives.[30] Now, equations (4.2) and (4.3) together imply once again that international prices ought to be used in benefit-cost analysis of public sector projects.[31] Using this fact we can rewrite equation (4.1) as

$$\frac{dV}{dq_i} + \lambda \left(\sum_{k \in A} \frac{\partial y_k}{\partial q_i} - \sum_{k \in A} \frac{dc_k}{dq_i} \right) = 0 \qquad i \in A \text{ and } i \neq 0, \tag{4.1a}$$

where

$$y_k = \sum_{j=1}^{m} y_k^j$$

The interpretation of equation (4.1a) is as before: the marginal cost in welfare per unit gain in the balance of payments should be the same regardless of which price (tariff) is varied. Now, since the government can be viewed as controlling both trade and public production, and since the shadow prices in the public sector are equal to the international prices, we can treat trade flow and public production indifferently. Writing $e_k = c_k - y_k = x_k + z_k$ as the "excess demand" or "imports" (from public production and foreign trade)[32] for commodity k we obtain from equation (4.1a) the fact that

---

[29] The actual controls are tariffs, t, and government production x. See above, p. 10.

[30] That is,

$dV/dq_i = (\partial V/\partial q_i) + [(\partial \pi/\partial q_i)(\partial V/\partial M)]$, $dc_k/dq_i = (\partial c_k/\partial q_i) + [(\partial \pi/\partial q_i)(\partial c_k/\partial M)]$

[31] The evaluation of projects in the private sector follows as in the previous section.

$$\sum_{k \in A} \frac{t_k}{e_i} \left(\frac{\partial e_i}{\partial q_k} \right)_{\bar{U}} = -\theta,^{33} \tag{4.4}$$

where

$$\theta = 1 - \frac{\partial V / \partial M}{\lambda} - \sum_{k \in A} t_k \frac{\partial c_k}{\partial M}$$

That is to say, the percentage reduction in "excess demand" should be the same for all commodities

Now clearly

$$\eta^e_{ik} \equiv - \left(\frac{\partial \ln e_i}{\partial \ln q_k} \right)_{\bar{U}} = \eta^d_{ik} + \frac{y_i}{e_i} (\eta^s_{ik} + \eta^d_{ik})$$

Therefore, equation (4.4) can be rewritten (on writing $\hat{t}_i = t_i/q_i = t_i/1 + t_i$) as

$$\hat{t}\ [\eta^e] = \hat{t} \left[[\eta^d] + \left(\frac{y}{e} \right) [I][\eta^s + \eta^d] \right] = \theta E \tag{4.5}$$

(where I is the identity matrix and $E \equiv (1, 1, 1, \quad , 1))$. In equation (4.5) (y/e) denotes a vector an element of which is (y_k/e_k). We can re-express equation (4.5) to obtain the optimal tariff structure as

$$\hat{t} = \theta E\ [\eta^e]^{-1} = \theta E\ [[\eta^d] + (y/e)\ [\eta^s + \eta^d]]^{-1} \tag{4.6}$$

We should contrast equation (4.5) with the results of section 3.2, as embodied in equation (3.7)

---

[32] We shall use the quotation marks on "excess demand" or "imports" to signify this

[33] Equation (4.1a) can be rewritten as

$$\frac{\partial V}{\partial q_i} + y_i \frac{\partial V}{\partial M} + \lambda \sum_k \frac{\partial y_k}{\partial q_i} - \lambda \sum_k \frac{\partial c_k}{\partial q_i} - \lambda y_i \sum_k \frac{\partial c_k}{\partial M} = 0 \qquad i \in A,$$

or

$$-(x_i + z_i) \frac{\partial V}{\partial M} + \lambda \sum_k (q_k - t_k) \frac{\partial y_k}{\partial q_i} - \lambda \sum_k (q_k - t_k) \frac{\partial c_k}{\partial q_i} - \lambda y_i \sum \frac{\partial c_k}{\partial M} = 0,$$

or

$$\begin{aligned} -\lambda \sum_{k \in A} t_k \left[\frac{\partial y_k}{\partial q_i} - \left(\frac{\partial c_k}{\partial q_i} \right)_{\bar{U}} \right] &= (x_i + z_i) \frac{\partial V}{\partial M} - \lambda c_i + \lambda \sum_{k \in A} t_k c_i \frac{\partial c_k}{\partial M} \\ &\quad + \lambda y_i \sum_{k \in A} (q_k - t_k) \frac{\partial c_k}{\partial M} \\ &= (x_i + z_i) \left(\frac{\partial V}{\partial M} + \lambda \sum_{k \in A} t_k \frac{\partial c_k}{\partial M} - \lambda \right), \end{aligned}$$

r, if $e_i \neq 0$,

$$\sum_{k \in A} \frac{t_k}{e_i} \left(\frac{\partial e_i}{\partial q_k} \right)_{\bar{U}} = \sum_{k \in A} t_k \frac{\partial c_k}{\partial M} - 1 + \frac{1}{\lambda} \frac{\partial V}{\partial M} = -\theta$$

In the case of independent demand and supply curves, equation (4 6) reduces readily to

$$\hat{t}_i = \frac{1}{\eta^e_{ii}} = \frac{\theta}{\eta^d_{ii} + \frac{y_i}{e_i}(\eta^d_{ii} + \eta^s_{ii})} \qquad i \in A. \tag{4 7}$$

The formula given in (4 7) expresses the fact that, at the optimum, the tax rate (as a percentage of the consumer price) is inversely proportional to the elasticity of "excess demand " Notice also that equation (4.7) implies an export tax for exports whose net supply curve is positively sloped, and a subsidy for exports whose net supply curve is backward bending It is clear then that duties need not be uniform in this second-best revenue tariff [34]

We turn now to some other cases For instance, if $y_i = 0$, which implies that the commodity does not enter into domestic production, then the tariff is identical to that described in section 3 2, as embodied in equation (3 7)

For intermediate goods, $y_i = -e_i$, so the tariffs/taxes are chosen to make the producers prices the same as discussed in section 3 2 (cf eq [3 7]) Note that this implies a tariff structure with different rates on different commodities, but such that the production of all intermediate goods is increased the same percentage, imports of intermediate goods not produced domestically are reduced by the same percentage

But not even this simplicity of structure obtains if firms are foreign owned For foreign-owned firms, equation (4 4) must be modified to read[35]

$$\sum_{k \in A} \frac{t_k}{e_i}\left(\frac{\partial e_i}{\partial q_k}\right)_U = -\theta + \frac{y_i}{e_i}\left(\frac{\partial V}{\partial M} \cdot \frac{1}{\lambda} + \sum t_k \frac{\partial c_k}{\partial M}\right) \tag{4 8}$$

Since raising the tariff increases profits, which accrue in this case to foreigners and hence hurt the balance of payments, tariffs on commodities produced by foreign-owned firms should be lower than on those produced by domestically owned firms (for the same elasticities of the excess demand

---

[34] See Ramaswami and Srinivasan (1968)

[35] Equation (4 1a) becomes

$$\frac{\partial V}{\partial q_i} + \lambda \sum \left(\frac{\partial y_k}{\partial q_i} - \frac{\partial c_k}{\partial q_i}\right) - \lambda y_i = 0,$$

or

$$-c_i \frac{\partial V}{\partial M} - \lambda \sum (q_k - t_k)\left(\frac{\partial e_k}{\partial q_i}\right) - \lambda y_i = 0$$

Hence,

$$\lambda \sum t_k \left(\frac{\partial e_k}{\partial q_i}\right)_{\bar{U}} = \lambda\left[c_i \sum_k \left(\frac{\partial c_k}{\partial M}\right) t_k - c_i + y_i\right] + c_i \frac{\partial V}{\partial M}$$

$$= \lambda\left[-e_i + e_i \sum \frac{\partial c_k}{\partial M} t_k + y_i \sum t_k \frac{\partial c_k}{\partial M}\right] + (e_i + y_i)\frac{\partial V}{\partial M}$$

curves), the difference in the percentage reduction in "imports" from what they would have been without tariffs is proportional to the ratio of domestic production to "imports."

4 2 Possibility of Taxing Profits at 100 Percent

There is one further case that deserves brief mention. We now suppose that the government can (and therefore, does) set the tax on pure profits at 100 percent Then, (4 1) becomes

$$\sum_k \frac{t_k}{e_i}\left(\frac{\partial e_i}{\partial q_k}\right)_{\bar{U}} = -\theta \frac{c_i}{e_i} \tag{4.9}$$ [36]

This implies that the relative reduction in "excess demand" ("imports") is simply inversely proportional to the ratio of "excess demand" ("imports") to consumption In particular, in the case of intermediates used to produce a nonconsumed export good, note that no tariffs should be imposed on imports Thus, the conclusion of Ramaswami and Srinivasan (1968), that "inputs used in export production must be free of duty" is only partially correct [37] if the tax authorities can impose only import duties and 100 percent profits taxes, then only if the good is also not consumed should it be exempt from duties If it can impose 100 percent profits taxes and consumption-production taxes as well as tariffs, no intermediate, whether used for producing exports or import substitutes, should be taxed If profits are not taxed at 100 percent, and if the same tax rate must be imposed on domestic production as on imported intermediates, then, even though the good is not consumed, a tariff-production tax ought to be imposed

5. Benefit-Cost Analysis with Given Taxes

5 1 All Taxes Fixed

We now consider the case where the government project evaluator needs to take the tariff and tax structure as *given* and *fixed*. The question is what shadow prices ought he to use in project evaluation? Notice that the number of controls in the planning exercise is now drastically reduced The Lagrangian of the planning problem is still expression (3 1), but now the set of controls to be chosen are the vector of public production, x,

---

[36] The introduction of nontradeables does not change these results Letting $\Gamma_k \equiv (p_k/\lambda) - p_k$, $k \in N$ be the difference between the shadow price and the private producer price of a nontradeable, (4 9) can be rewritten

$$\sum_{k\in T} t_k \left(\frac{\partial e_i}{\partial q_k}\right)_{\bar{U}} + \sum_{k\in N} \Gamma_k \left(\frac{\partial e_i}{\partial q_k}\right) = -\theta c_i$$

[37] Compare Ramaswami and Srinivasan (1968, p 576)

and the volumes of trade, z. If all goods were traded, fixing all taxes would fix all consumer and producer prices. Since our focus in this section is explicitly on public production, it is convenient to assume there is a single public good; resource savings from greater efficiency in the public sector can be used to increase the expenditure on the public good. Letting g denote the public good, our Lagrangian can now be formulated:[38]

$$\mathcal{L} = V(q, g) - \lambda \sum_{k\in T} z_k + \sum_{k\in A} \rho_k(x_k + z_k + y_k - c_k) + \mu G(x, g)$$

The first-order conditions now read as

$$V_g + \mu G_g - \sum \rho_k \frac{\partial c_k}{\partial g} = 0, \qquad (5.1)$$

$$-\lambda + \rho_k = 0 \qquad k \in T, \qquad (5.2)$$

$$\rho_k + \mu \frac{\partial G}{\partial x_k} = 0 \qquad k \in D. \qquad (5.3)$$

From equations (5.2) and (5.3), we note that for traded goods the marginal rates of transformation in the *public* sector must equal the international price ratios, unity. That is, the project evaluator ought to use international prices in project evaluation. Notice that we have not assumed that the fixed tax rates are in any sense optimal. For nontradeables, we note a result similar to that obtained in section 3.3: the shadow price of a nontraded good is its marginal foreign exchange cost. (It is obvious that this set of results continues to hold if profits can be taxed up to a 100 percent.)

Turning to the "evaluation" of a private domestically owned project, the analysis for arbitrary but fixed tariffs and taxes is identical to that of section 3 where it was supposed that tariffs and taxes were optimally chosen.

5.2 *Some Taxes Controllable*

How does the fact that *some* tariff rates are unalterable affect the choice of the remaining tariff rates? Assume that, for one set of commodities, say R, q_i can be chosen by the government but that, for the remaining goods, they are given as fixed. Assume for notational ease that all goods are tradeable, that is, $A = T$. It is then trivial to confirm, on repeating the analysis of section 3.1, that

$$\sum_{k\in R} t_k \left(\frac{\partial c_i}{\partial q_k}\right)_U + \sum_{k\notin R} t_k \left(\frac{\partial c_i}{\partial q_k}\right)_U = -c_i\theta \qquad i \in R \qquad (5.4)$$

---

[38] We assume for notational simplicity in this formulation that the public good is not traded, and all other goods are. The reader can easily extend the analysis for the situation where there are nontraded private goods in the economy.

It follows that the consumption of all freely taxable goods should be reduced by the same percentage. This implies that, if demand elasticities are assumed independent, the tax rates on those commodities for which we can choose tax rates are unaffected.

5.3 *Quotas*

It is not surprising that the presence of quotas alters both the tariff structure of the remaining commodities and the shadow prices to be used in the public sector for those commodities on which quotas have been imposed. To the basic Lagrangian (3.1) we now add an extra term to obtain

$$\mathcal{L} = V\left[q, \sum_j \pi^j(p^j)\right] - \lambda \sum_{k \in T} z_k + \mu G(x) + \sum_{j=1}^{m} \psi_j F^j(y^j)$$
$$+ \sum_{k \in A} \rho_k \left(\sum_j y_k^j + x_k + z_k - c_k \right) + \gamma(\bar{z}_s - z_s), \qquad (5.5)$$

where $\bar{z}_s$ is the quota imposed on commodity s. It is immediate that the first-order conditions (3.4) and (3.5) continue to hold here for all $i \in D$ and $i \neq s$. If s enters into domestic production, then the first-order conditions pertaining to its shadow price in the public sector read

$$\mu \frac{\partial G}{\partial x_s} + \rho_s = 0, \qquad (5.6)$$

$$-\lambda + \rho_s - \gamma = 0 \quad \text{where } \gamma \geq 0 \text{ and } \gamma(\bar{z}_s - z_s) = 0, \qquad (5.7)$$

from which we obtain the fact that $(\partial G/\partial x_s)/(\partial G/\partial x_0) = 1 + \gamma/\lambda \geq 1$. Generally speaking then, the shadow price of s in the public sector is equal to unity only when the quota does not bite. Otherwise it is greater than unity. Turning to the structure of optimum taxes we note simply that both the tariffs and the production taxes are affected. For instance, if we assume that tariffs may be chosen on all commodities and that s is consumable, then the equation (corresponding to eq. [3.6]) yielding the structure of taxes on consumer goods reads as

$$\sum_{k \in C} \frac{t_k}{c_i} \left(\frac{\partial c_i}{\partial q_k} \right)_{\bar{v}} - \frac{\gamma_s}{\lambda c_i} \left(\frac{\partial c_i}{\partial q_s} \right)_{\bar{v}} = -\theta \qquad i \in C \qquad (5.8)$$

In the case of independent demand curves, we notice from equation (5.8) that the tariff of only commodity s is altered, and it is set so as to capture the profits which would otherwise accrue to the importers of s. Complements of s have their consumption reduced proportionately more than is the consumption of substitutes. That is to say, the tax structure ought to be designed so as to encourage the consumption of substitutes for the commodity on which a quota has been imposed. This is clearly what intuition suggests.

6. Trade Policy Dependent on Project Selection

Quite often a decision maker realizes that the presence of a particular project in the country is likely to give rise to changes in the tariff policy. In particular, there is a long experience of newly established industries requesting and obtaining protection, even when on "purely economic" grounds this protection is questionable. To attempt to capture this kind of situation we postulate for simplicity of notation that for some commodity ℓ the tariff t_ℓ is an increasing function of the level of imports of ℓ That is, $t_\ell = t_\ell(z_\ell)$, $t'_\ell > 0$ Assume also for simplicity of notation that the model is otherwise the same as that in section 3. The drastic assumption that we make here, as the reader will immediately recognize, is the assuming away of any game theoretic problems that this kind of tariff "response" obviously has built within it We are thus assuming that the tariff on ℓ responds passively to the level of imports of ℓ. It is then simple to see that the equations corresponding to equations (3 4) and (3.5) read here as

$$\mu \frac{\partial G}{\partial x_i} + \rho_i = 0 \qquad i \in D, \tag{6 1}$$

$$-\lambda + \rho_i = 0 \qquad i \neq \ell \text{ and } i \in T, \tag{6 2}$$

$$\frac{\partial V}{\partial q_\ell}\frac{\partial t_\ell}{\partial z_\ell} - \lambda + \rho_\ell - \sum_{k \in C} \rho_k \frac{\partial c_k}{\partial q_\ell}\frac{\partial t_\ell}{\partial z_\ell} = 0. \tag{6 3}$$

It follows that, for all tradeable goods except ℓ, shadow prices in the public sector are their international prices. So far as commodity ℓ is concerned, we can reexpress equation (6 3) as

$$\rho_\ell = \lambda + \frac{\partial t_\ell}{\partial z_\ell}\left(c_\ell \frac{\partial V}{\partial M} + \sum_{k \in C} \rho_k \frac{\partial c_k}{\partial q_\ell}\right) \tag{6 4}$$

From equation (6.4), we can readily verify that the shadow price of commodity ℓ is[39]

$$1 - c_\ell \frac{\partial t_\ell}{\partial z_\ell}\left[\theta + \sum_{i \in C} \frac{t_i}{c_\ell}\left(\frac{\partial c_\ell}{\partial q_i}\right)_{\bar{U}}\right] \Big/ 1 - \frac{\partial c_\ell}{\partial q_\ell}\frac{\partial t_\ell}{\partial z_\ell}. \tag{6.5}$$

---

[39] $$\frac{\partial V/\partial M}{\lambda} + \sum \frac{\rho_k}{\lambda}\frac{\partial c_k}{\partial q_\ell}\frac{1}{c_\ell} = \frac{\partial V/\partial M}{\lambda} + \frac{1}{c_\ell}\sum\left(\frac{\rho_k}{\lambda} - q_k\right)\frac{\partial c_k}{\partial q_\ell} - 1$$
$$= \frac{\partial V/\partial M}{\lambda} - \frac{1}{c_\ell}\sum t_k \frac{\partial c_k}{\partial q_\ell} - 1 + \left(\frac{\rho_\ell}{\lambda} - 1\right)\frac{\partial c_\ell}{\partial q_\ell}\frac{1}{c_\ell}$$
$$= \frac{\partial V/\partial M}{\lambda} + \sum t_k \frac{\partial c_k}{\partial M} - 1 - \frac{1}{c_\ell}\sum\left(\frac{\partial c_\ell}{\partial q_k}\right)_{\bar{U}} t_k$$
$$+ \left(\frac{\rho_\ell}{\lambda} - 1\right)\frac{\partial c_\ell}{\partial q_\ell}\frac{1}{c_\ell}.$$

This analysis assumes all goods are tradeable The modifications required if some goods are not traded are obvious.

It follows from expression (6.5) that the shadow price of commodity ℓ is below or above the international price of ℓ depending on whether the given tax structure reduces its consumption by less or more than is optimal. In particular, if demand curves are independent it depends simply on whether $t_\ell/(1 + t_\ell) \gtrless t_\ell^*/(1 + t_\ell^*) \equiv \theta/\eta_{\ell\ell}^d$

More generally, it is simple to demonstrate (Dasgupta and Stiglitz 1971) that, for tradeable commodities, if the tariff on some commodity k depends on the level of imports of some other commodity k' as well as on, say, the government's net production of yet some other commodity, k'', then the shadow prices of *only* k' and k'' differ from their international prices It follows from this that the fact that the tariff on commodity k responds unoptimally to the import (or the public production) of some other commodity k' is no argument for setting the shadow price of k (or for that matter any commodity other than k') different from its international price We have found it surprising that this should be so in such an interdependent system as we have been considering

If demand curves are interdependent, then the tariffs on other goods will, however, be affected by the constraint on the tariff on commodity ℓ Just as in the case of quotas, substitutes for commodity ℓ will have their production reduced by less than complements

7. Foreign Exchange Constraint

Although the most natural interpretation of the models presented thus far is in terms of the conventional models of static international trade, there is no reason why the different commodities could not be treated as well as *dated* commodities As we noted earlier in section 1 2, a foreign exchange constraint may be interpreted as a limitation on the amount a country can borrow in any period To make the interpretation clearer, we let $c_{it}, y_{it}, x_{it}, z_{it}$ be, respectively, the consumption, net output in the private sector, net output in the public sector, and net imports of commodity i at time t We assume that all relative international prices (*within* each period) are constant, so that we normalize all at unity Let r_t be the international rate of interest at t[40] Then the trade balance condition can be written as

$$\sum_t \frac{\sum_i z_{it}}{1 + r_t} = 0 \tag{7.1}$$

The borrowing constraint at some year t' reads as

$$\sum_i z_{it'} \leq \varepsilon_{t'}. \tag{7.2}$$

---

[40] That is, $1/1 + r_t$ is the price today of a promise to deliver $1 00 of foreign exchange at time t.

The basic Lagrangian (3.1) now has an extra term, and it reads as

$$V\left[q, \sum_j \pi^j(p^j)\right] = \lambda \sum_t \frac{\sum_k z_{kt}}{1 + r_t} + \mu G(x) + \sum_{j=1}^{m} \psi_j F^j(y^j)$$

$$+ \sum_t \sum_k \rho_{kt}\left(\sum_{j=1}^{m} y^j_{kt} + x_{kt} + z_{kt} - c_{kt}\right)$$

$$+ \sum_{t'} \gamma_{t'}\left(\varepsilon_{t'} - \sum_k z_{kt'}\right), \tag{7.3}$$

where $\gamma_{t'} \geq 0$ and

$$\gamma_{t'}\left(\varepsilon_{t'} - \sum_k z_{kt'}\right) \geq 0.$$

The first-order conditions, corresponding to equations (3.4) and (3.5), now read as

$$\mu \frac{\partial G}{\partial x_{it}} + \rho_{it} = 0, \tag{7.4}$$

$$\rho_{it} - \frac{\lambda}{1 + r_t} = 0 \qquad t \neq t', \quad i \in T, \tag{7.5}$$

$$\rho_{it'} - \gamma_{t'} - \frac{\lambda}{1 + r_{t'}} = 0 \qquad i \in T. \tag{7.6}$$

Equations (7.4–7.6) yield the important implication that *within* any period international prices ought to be used in the evaluation of public sector projects. But if the constraint on foreign exchange is binding at t', the rate of interest (the rate of discount) at t' should not be $r_{t'}$; rather, it should be somewhat higher than that.[41] On the other hand, for nontraded goods, the shadow price is simply the value at international prices of the foreign exchange that could have been generated if the production of the nontraded good had been decreased by a unit and the production of traded goods increased.

If profits are taxed at 100 percent, then it is easy to confirm that the private sector ought to face the same set of prices as the public sector. That is to say, at the optimum the economy is productively efficient. With less than 100 percent profit tax this is no longer true, and indeed the rate of discount that should be used for different nontraded goods may well be different.

Turning to the structure of taxes, we note once again that consumption of commodities in periods which are substitutes for consumption in periods in which the constraint is binding is reduced by less than the consumption

[41] $(\partial G/\partial x_{it})/(\partial G/\partial x_{jt}) = \rho_{it}/\rho_{jt} = 1$; $(\partial G/\partial x_{it})/(\partial G/\partial x_{it'}) = \rho_{it}/\rho_{it'} = \lambda/(1 + r_t)/\{[\lambda/(1 + r_{t'})] + \gamma_{t'}\}$.

in periods that are complements. That is to say, if the constraint is binding for initial years, the tariff structure should be designed so as to encourage savings in the initial periods (e.g., by having higher taxes on complements of leisure in order to encourage work in initial periods) and a gradual lowering of tariffs over time (the substitution effect encouraging the postponement of consumption). We do not reproduce the formal argument here, since it is pretty straightforward.

8. Budget Constraint

In this section we consider a constraint on the government budget deficit in any given year. For a closed economy, this constraint was looked at originally by Boiteux (1956), and his analysis was extended in Stiglitz and Dasgupta (1971).

The constraint may be written as[42]

$$\sum_{i\in D} q_{it}x_{it} \geq b_t. \tag{8.1}$$

If δ_t is the dual associated with the constraint (8.1) then we have, on writing

$$\hat{\lambda}_t \equiv \frac{\lambda}{\prod_{n=1}^{t}(1+r_n)},$$

the simple result

$$\frac{\partial G/\partial x_{it}}{\partial G/\partial x_{kt}} = \frac{\hat{\lambda}_t + \delta_t q_{it}}{\hat{\lambda}_t + \delta_t q_{kt}} \qquad i, k \in D \cap T \tag{8.2}$$

Without loss of generality, let $q_{it} > q_{kt}$. Then

$$\frac{q_{it}}{q_{kt}} > \frac{\partial G/\partial x_{it}}{\partial G/\partial x_{kt}} > 1 \tag{8.3}$$

The shadow prices to be used in the public sector lie between the international prices and the domestic prices, regardless of whether the domestic prices are determined by optimal or nonoptimal tariffs and taxes. Similarly,

$$\frac{\partial G/\partial x_{it}}{\partial G/\partial x_{it+1}} = \frac{\hat{\lambda}_t + \delta_t q_{it}}{(\hat{\lambda}_t/1 + r_{t+1}) + \delta_{t+1}q_{it+1}}. \tag{8.4}$$

---

[42] In writing (8.1), we have implicitly assumed that $D = C$; for intermediate goods, we have to use production prices. This does not change the results at all.

If we were to assume $q_{it} = q_{it+1}$, then

$$\frac{\partial G/\partial x_{it}}{\partial G/\partial x_{it+1}} \gtrless (1 + r_{t+1}) \text{ as } \frac{\delta_{t+1}}{\delta_t} \lesseqgtr \frac{1}{1 + r_{t+1}} \tag{8.5}$$

Condition (8.5) gives a qualitative rule for choice of the rate of discount in the public sector.[43]

9. Distributional Objectives

Thus far we have been assuming the existence of a representative individual whose welfare the government is maximizing. It is plain, however, that considerations of income distribution ought to weigh heavily both in the determination of the structure of taxes and tariffs and in the selection of investment projects in the public sector. Assume that the government is interested in maximizing the individualistic social welfare function of the form $W = W(U_1, \ldots, U_r)$ where there are r nonidentical individuals in the economy. It is then easy to show that corresponding to W there is an indirect social welfare function

$$V = V(q, M_1, \ldots, M_r) \tag{9.1}$$

where, if β_j^k is the share of the jth firm owned by individual k,

$$M_k = \sum_j \beta_j^k \pi^j$$

Then, in the analysis of sections 1–8 we simply replace $V[q, (1 - \tau_\pi)\pi, E]$ by the function defined in expression (9.1). It is then easy to show (see Dasgupta and Stiglitz 1971) that, in this instance, even though the structure of optimal taxes and tariffs is affected by distributional considerations, the basic qualitative propositions concerning public investment criteria (i.e., the nature of the shadow prices to be used) remain unchanged. That is to say, the government will still use international prices to evaluate public projects in those cases where, when we ignored distributional objectives, it was optimal to do so.[44]

10. Conclusions

We now summarize the basic results obtained in this paper in the form of a few simple rules.

---

[43] It is, perhaps, of some interest to note that if profits were taxed at 100 percent and the only constraint facing the government were the budget constraint (8.1), the optimum tax structure would be such that the prices of tradeable goods faced by the *private* sector would be unity. That is to say, at the optimum $\tau_i = 0$ for all i (see Dasgupta and Stiglitz 1971).

[44] For a fairly extensive discussion of the implications of distributional objectives for the structure of taxation in a closed economy, see Atkinson and Stiglitz (1973).

10.1 Public Investment Criteria

Rule 1. The public sector ought to use a uniform set of shadow prices in all its projects.[45]

Rule 2. The shadow price of a tradeable commodity is its international price unless (*a*) there is a government budgetary constraint, (*b*) there is a foreign exchange constraint, (*c*) there is a quota on that commodity, or (*d*) the level of net imports (or net *public* production) of that particular commodity influences unoptimally some control (e.g., some tariff) at the disposal of the government.[46]

Rule 3. When there is a government budgetary constraint, the shadow price of tradeable lies between the world price and its domestic price.

Rule 4. When there is a foreign exchange constraint the relative prices of commodities within each period are equal to the world price ratios; but the rate of interest used in project evaluation is not equal to the world rate of interest.

Rule 5. When there is a quota on a particular commodity, its shadow price should be greater than its international price, if the quota is binding.

Rule 6. When the net level of imports of a commodity influences its tariff level, its shadow price should be greater or less than its international price depending on whether as a consequence of the tariff response the consumption of the commodity is reduced (from what it would have been had the tariff been directly controllable and had been at its optimal level) or increased.

Rule 7. Except under those exceptions noted in rule 2, nontradeables ought to be valued at their "foreign exchange" equivalent. That is the value of the foreign exchange that would be earned if one less unit of the given nontradeable were produced and the resources diverted to the production of tradeables.

What is particularly useful about these results is that they show the project evaluator what to do even in "second-best" (optimally chosen tariffs and taxes) or "third-best" (nonoptimally chosen tariffs and taxes) situations. They show that, even when there are quotas (or similar restrictions), although one does not use international prices for the commodities on which there is a quota, for other commodities one does.

10.2. Tariffs and Taxes

All the following rules are predicated on the assumption that lump-sum taxation is unfeasible, so that the government has to resort to distortionary taxes to raise revenue. These rules are not dependent on the assumption of a representative consumer. (Although our analysis employed this assumption, a careful reading of the proofs show that it nowhere entered the arguments for any of these results, or any of the results cited above.)

---

[45] This does not, of course, pertain to projects at different locations in an economy where there are substantial transport costs. As in usual general equilibrium analysis, one would then wish to expand the commodity space to account for locational differences.

[46] There is a sense in which (*c*) may be viewed as a special case of (*d*). A quota is like a discontinuous tariff imposition.

Rule 1. In a centrally controlled economy, or in an economy where 100 percent profit taxes may be levied, and in which consumption, trade, and production taxes may be imposed, no taxes (either on trade or on domestic production) should be levied on intermediates Only general consumption taxes ought to be imposed

Rule 2. Under the same circumstances as in rule 1, if the only taxes which can be levied are trade taxes, then the output of intermediate goods should not be changed from what it would be at international prices. Goods which are used both as inputs into production and as consumption goods should be taxed (if it is impossible to treat the same good differently according to use)

Rule 3 In an economy where profits are not completely taxed away, both consumption and trade taxes should be employed. Imported intermediate goods should not be taxed if they can be treated differently from the (same) domestically produced intermediate goods; otherwise they should be

Rule 4 If profits are not taxed at 100 percent, and firms are foreign owned, then the production tax ought to be such as to reduce the output of all commodities by the same percentage; the tax is independent of the desire of the government for tax revenue (see eq [3 16])

The following rules provide the detailed form of the tax and tariff structure (for small taxes) under the assumption of a "representative" consumer The exact formulae that are valid regardless of the size of the revenue are given by the equations in parentheses in the text

Rule 5 If there is no budget constraint, no foreign exchange constraint, no quotas, and no commodity whose tariff responds unoptimally to the level of import or public production, and if it is feasible to impose both consumption and production taxes,[47] then the tariffs and export duties should be such that the consumption of all commodities is reduced (along the compensated demand curve) by the same percentage from what it would have been had international prices been charged (eqq [2 18] and [3 6])

Rule 6 If profits are not taxed at all and all firms are domestically owned, then there should be a tax on domestic production such that the output of all commodities is reduced exactly by the same percentage as that of the consumption of all commodities due to the trade taxes[48] (see eq [3.8]).

---

[47] The same qualifications apply to rules 6 and 7.

[48] As we noted earlier, intermediate goods are treated just like final goods. Recall that we are treating outputs and inputs symmetrically.

Rule 7 If only trade can be taxed and there is no profit tax, then the trade taxes ought to be chosen so as to reduce "excess demand" of each commodity by the same percentage (again along the compensated demand curve) from what it would have been had international prices been charged (eq [4 4]) If firms are foreign owned, the relative reduction in excess demand should be smaller the smaller is the ratio of excess demand to domestic private production (eq [4 8]) If there is 100 percent profit taxation, then the reduction in excess demand should be inversely proportional to the ratio of excess demand to consumption[49] (eq [4 8a])

Rule 8 If some tariffs (export duties) are fixed, then the remaining duties ought to be chosen with a view to seeing that the consumption of commodities for which trade taxes can be varied is reduced by the same percentage (along the compensated demand curve) from what it would have been had international prices for all commodities been charged.[50]

Rule 9 If there is a quota on commodity *k* or if the level of tariff on *k* is an increasing function of the net import of *k*, then the tariffs are chosen so as to ensure that the consumption of complements of *k* is reduced by more than is the consumption of substitutes

The rules for the evaluation of projects in the private sector are considerably more complicated Typical of the kind of results available is the following, for the case where trade taxes are given and profits are untaxed, the profits, valued at *domestic* prices must be greater than the foreign exchange cost (valued at international prices) times the shadow price of foreign exchange in terms of domestic consumption

References

Atkinson, A B, and Stiglitz, J E "The Structure of Indirect Taxation and Economic Efficiency." *J Public Econ* 1 (April 1972) 97-119

——— "Alternative Approaches to the Redistribution of Income" Mimeographed 1973

Bhagwati, J N "Generalized Theory of Distortions and Welfare" In *Trade, Balance of Payments and Growth*, edited by J N Bhagwati, R W Jones, R Mundell, and J Vanek Amsterdam North-Holland, 1971

Boiteux, M "Sur la gestion des monoples publics astreints a l'equilibre budgetaire" *Econometrica* 24 (January 1956) 22-40

Corden, W M "The Structure of a Tariff System and the Effective Protective Rate" *J P E* 74 (June 1966) 221-37

Dasgupta, P, Marglin, S A, and Sen, A K *Guidelines for Project Evaluation*. New York United Nations, 1972

Dasgupta, P, and Stiglitz, J E "Valuation of Tradeables in Benefit-Cost Analysis." Mimeographed 1971

---

[49] In particular, this means that output of intermediate goods is not reduced at all.

[50] Rules 8 and 9 are predicated on there being both consumption (trade) and production taxes for the commodities in question Moreover, we assume that profits are not taxed

———. "On Optimal Taxation and Public Production." *Rev. Econ. Studies* 39 (January 1972): 87–104.

Debreu, G. *The Theory of Value*. New York: Wiley, 1959.

Diamond, P. A., and Mirrlees, J. A. "Optimal Taxation and Public Production." *A.E.R.* 61 (March-June 1971): 8–27, 261–78.

Little, I M. D., and Mirrlees, J. A. *Manual of Industrial Project Analysis in Developing Countries*. Paris: Org Econ. Cooperation and Development, 1969.

Marglin, S. A. "Value and Price in the Labour Surplus Economy." Mimeographed Cambridge, Mass.: Harvard Univ., 1972.

Meade, J. E. *Trade and Welfare*. Oxford · Oxford Univ. Press, 1955.

Mirrlees, J. A. "On Producer Taxation." *Rev Econ. Studies* 39 (January 1972): 105–12.

Pigou, A. C. *A Study in Public Finance*. New York · Macmillan, 1947.

Ramaswami, V. K., and Srinivasan, T. N. "Optimal Subsidies and Taxes When Some Factors Are Traded." *J.P E* 76 (July/August 1968): 569–82.

Ramsey, F. P. "A Contribution to the Theory of Taxation " *Econ. J* 37 (March 1927): 47–61.

Ruffin, R. J. "Tariffs, Intermediate Goods, and Domestic Protection." *A.E.R.* 59 (January 1969) 29–69

Sen, A K. "Control Areas and Accounting Prices · An Approach to Project Evaluation." *Econ. J.* 82 (March 1972) · 486–501.

Stiglitz, J. E "Alternative Theories of Wage Determination and Unemployment in L.D.C.'s. I The Labor Turnover Model " *Q J.E* (in press).

———. "Alternative Theories of Wage Determination and Unemployment. II The Efficiency Wage Hypothesis " Discussion Paper no. 357, Cowles Found., Yale Univ , 1973

Stiglitz, J. E., and Dasgupta, P "Differential Taxation, Public Goods, and Economic Efficiency." *Rev. Econ. Studies* 38 (April 1971): 151–74.

Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition

Sherwin Rosen

University of Rochester and Harvard University

A class of differentiated products is completely described by a vector of objectively measured characteristics Observed product prices and the specific amounts of characteristics associated with each good define a set of implicit or "hedonic" prices A theory of hedonic prices is formulated as a problem in the economics of spatial equilibrium in which the entire set of implicit prices guides both consumer and producer locational decisions in characteristics space Buyer and seller choices, as well as the meaning and nature of market equilibrium, are analyzed. Empirical implications for hedonic price regressions and index number construction are pointed out

I. Introduction and Summary

This paper sketches a model of product differentiation based on the hedonic hypothesis that goods are valued for their utility-bearing attributes or characteristics Hedonic prices are defined as the implicit prices of attributes and are revealed to economic agents from observed prices of differentiated products and the specific amounts of characteristics associated with them They constitute the empirical magnitudes explained by the model Econometrically, implicit prices are estimated by the first-step regression analysis (product price regressed on characteristics) in the construction of hedonic price indexes With few exceptions, structural

---

The substance of this paper arose from conversations with H Gregg Lewis several years ago A multitude of other people have contributed advice and criticism Among them are William Brock, Stanley Engerman, Robert J Gordon, Zvi Griliches, Robert E Lucas, Jr., Michael Mussa, and the referee Remaining errors are my own responsibility Financial support from the Center for Naval Analysis and the National Institute of Education is gratefully acknowledged

interpretations of the hedonic method are not available.[1] Therefore, our primary goal is to exhibit a generating mechanism for the observations in the competitive case and to use that structure to clarify the meaning and interpretation of estimated implicit prices. It will be shown that these data generally contain less information than is commonly supposed However, the model suggests a method that often can identify the underlying structural parameters of interest Also, as a general methodological point, it is demonstrated that conceptualizing the problem of product differentiation in terms of a few underlying characteristics instead of a large number of closely related generic goods leads to an analysis having much in common with the economics of spatial equilibrium and the theory of equalizing differences

The model itself amounts to a description of competitive equilibrium in a plane of several dimensions on which both buyers and sellers locate The class of goods under consideration is described by n objectively measured characteristics Thus, any location on the plane, is represented by a vector of coordinates $z = (z_1, z_2, \ \ , z_n)$, with z_i measuring the amount of the ith characteristic contained in each good Products in the class are completely described by numerical values of z and offer buyers distinct packages of characteristics Furthermore, existence of product differentiation implies that a wide variety of alternative packages are available Hence, transactions in products are equivalent to tied sales when thought of as bundles of characteristics, suggesting applicability of the principle of equal advantage for analyzing market equilibrium

In particular, a price $p(z) = p(z_1, z_2, \ \ , z_n)$ is defined at each point on the plane and guides both consumer and producer locational choices regarding packages of characteristics bought and sold Competition prevails because single agents add zero weight to the market and treat prices $p(z)$ as parametric to their decisions In fact the function $p(z)$ is identical with the set of hedonic prices—"equalizing differences"—as defined above, and is determined by some market clearing conditions· Amounts of commodities offered by sellers at every point on the plane must equal amounts demanded by consumers choosing to locate there Both consumers and producers base their locational and quantity decisions on maximizing behavior, and equilibrium prices are determined so that buyers and sellers are perfectly matched No individual can improve his position, and all optimum choices are feasible As usual, market clearing prices, $p(z)$, fundamentally are determined by the distributions of consumer tastes and producer costs We show how it is possible to recover, or identify, some of the parameters of these underlying distributions by a suitable transformation of the observations.

---

[1] Excellent summaries of the hedonic technique are available in Griliches (1971, chap. 1) and Gordon (1973). Major exceptions to the statement in the text are those studies dealing with depreciation and obsolescence (see Griliches 1971, chaps. 7 and 8) and some recent models based on markup pricing (e.g , Ohta and Griliches 1972)

An early contribution to the problem of quality variation and the theory of consumer behavior has been made by Houthakker (1952). His analysis is designed to take account of the fact that consumers purchase truly negligible fractions of all goods available to them without having to deal with a myriad of corner solutions required by conventional theory. That virtue of Houthakker's treatment is preserved in the present model. More recently Becker (1965), Lancaster (1966), and Muth (1966) have extended Houthakker's methods to more explicit consideration of utility-bearing characteristics. Again, the emphasis is on consumer behavior and properties of market equilibrium have not been worked out, a gap we hope to fill, in part, here. The spirit of these recent contributions is that consumers are also producers. Goods do not possess final consumption attributes but rather are purchased as inputs into self-production functions for ultimate characteristics. Consumers act as their own "middlemen," so to speak. In contrast, the model presented below interposes a *market* between buyers and sellers. Producers themselves tailor their goods to embody final characteristics desired by customers and receive returns for serving economic functions as intermediaries. These returns arise from economies of specialized production achieved by specialization and division of labor through market transactions not available outside organized markets with self-production.

Section II discusses individual choices in the market and the nature of market equilibrium. Some simple examples of analytic solutions for general equilibrium are given in Section III. Section IV presents an empirical method for identifying the underlying structure from the observations, while Section V applies the model to price index number construction in the presence of legislated restrictions. To highlight essential features, the simplest possible specifications are chosen throughout. As a further appeal to intuition, use is made of geometrical constructions wherever possible.

II. Market Equilibrium

Consider markets for a class of commodities that are described by n attributes or characteristics, $z = (z_1, z_2, \ldots, z_n)$. The components of z are objectively measured in the sense that all consumers' perceptions or readings of the amount of characteristics embodied in each good are identical, though of course consumers may differ in their subjective valuations of alternative packages. The terms "product," "model," "brand," and "design" are used interchangeably to designate commodities of given quality or specification. It is assumed that a sufficiently large number of differentiated products are available so that choice among various combinations of z is continuous for all practical purposes. That is, there is a "spectrum of products" among which choices can be made. As will be apparent, this assumption represents an enormous simplification of the problem. It is obviously better approximated in some markets than others, and there is no need to belabor its realism.[2] To avoid complications of capital theory, possibilities for resale of used items in secondhand markets are ignored, either by assuming that secondhand markets do not exist, or alternatively, that goods represent pure consumption.

Each product has a quoted market price and is also associated with a fixed value of the vector z, so that products markets implicitly reveal a function $p(z) = p(z_1, \ldots, z_n)$ relating prices and characteristics. This function is the buyer's (and seller's) equivalent of a hedonic price regression, obtained from shopping around and comparing prices of brands with different characteristics. It gives the minimum price of any package of characteristics. If two brands offer the same bundle, but sell for different prices, consumers only consider the less expensive one, and the identity of sellers is irrelevant to their purchase decisions. Adopt the convention of measuring each z_i so that they all may be treated as "goods" (i.e., so that consumers place positive rather than negative marginal valuations on them) in the neighborhood of their minimum technically feasible amounts. Then firms can alter their products and increase z only by use of additional resources, and $p(z_1, \ldots, z_n)$ must be increasing in all its arguments. Assume $p(z)$ possesses continuous second derivatives. Since a major goal of the analysis is to present a picture of how $p(z)$ is determined, it is inappropriate to place too many restrictions on it at the outset. However, note that there is no reason for it to be linear as is typically the case. The reason is that the differentiated products are sold in separate, though of course highly interrelated, markets. This point is spelled out in some detail below.

A buyer can force $p(z)$ to be linear if certain types of arbitrage activities are allowed. Let z_a, z_b, and z_c be particular values of the vector z. (i) Suppose $z_a = (1/t)z_b$, and $p(z_a) < (1/t)p(z_b)$, where t is a scalar and $t > 1$. Then t units of a model offering z_a yield the same amount of characteristics as a model offering z_b, but at less cost, ruling out transactions in convex portions of $p(z)$. (ii) Suppose $z_a < z_b < z_c$ and $p(z_b) > \delta p(z_a) + (1 - \delta)p(z_c)$, where $0 < \delta < 1$ and z_b is defined by $z_b = \delta z_a + (1 - \delta)z_c$. Then characteristics in amount of z_b could be achieved by purchasing δ units of a model containing z_a and $(1 - \delta)$ units of a model containing z_c at lower cost than by direct purchase of a brand containing z_b, and products in concave portions of $p(z)$ would be uneconomical. Arbitrage is assumed impossible in what follows (at this point we depart from Lancaster [1966]) on the assumption of indivisibility. This amounts to an assumption that packages cannot be untied. For example, in terms of one characteristic, two 6-foot cars are not equivalent to one 12 feet in length, since they cannot be driven simultaneously (case [i]), while a 12-foot car for half a year and a 6-foot car for the other half is not the same as 9 feet all year round (case [ii]). Similarly, assume sellers cannot repackage existing products in this manner or do not find it economical to do so, as might not be the case with perfect rental markets and zero transactions and reassembly costs.

---

[2] This assumption was first employed by L. M. Court (1941) and allows the use of marginal analysis rather than the programming methods required by Lancaster's (1966) formulation. Following the general rule, it is not without its costs, however (see below).

A. The Consumption Decision

To begin, suppose consumers purchase only one unit of a brand with a particular value of z. Write the utility function as $U(x, z_1, z_2, \ldots, z_n)$ assumed strictly concave, in addition to the other usual properties, where x is all other goods consumed. It would not be difficult to treat z as intermediate goods and relate them to yet more ultimate commodities through self-production functions, but that complication is ignored. Set the price of x equal to unity and measure income, y, in terms of units of x: $y = x + p(z)$. Maximization of utility subject to the nonlinear budget constraint requires choosing x and $(z_1, \ldots, z_n)$ to satisfy the budget and the first-order conditions $\partial p/\partial z_i = p_i = U_{z_i}/U_x$, $i = 1, \ldots, n$. Optimality is achieved by purchasing a brand offering the desired combination of characteristics. Second-order conditions are fulfilled on the usual assumptions regarding U, so long as $p(z)$ is not sufficiently concave (for a general statement of these conditions under a nonlinear constraint see Intriligator [1971]).

To stress the essential spatial context of the problem, define a value or bid function $\theta(z_1, \ldots, z_n, u, y)$ according to

$$U(y - \theta, z_1, \ldots, z_n) = u. \tag{1}$$

The expenditure a consumer is willing to pay for alternative values of $(z_1, \ldots, z_n)$ at a given utility index and income is represented by $\theta(z, u, y)$. It defines a family of indifference surfaces relating the z_i with "money" (i.e., with x foregone), and has been widely used in urban economics (e.g., see Alonso 1964). Differentiate (1) to obtain

$$\theta_{z_i} = U_{z_i}/U_x > 0,\ \theta_u = -1/U_x < 0, \text{ and } \theta_y = 1, \tag{2}$$

$$\theta_{z_i z_i} = (U_x^2 U_{z_i z_i} - 2U_x U_{z_i} U_{x z_i} + U_{z_i}^2 U_{xx})/U_x^3 < 0, \tag{3}$$

where the inequality in (3) follows from the assumptions about the bordered Hessian matrix of U. Also, strict concavity of U implies that θ is concave in z. Equations (2) and (3) show that the value function is increasing in z_i at a decreasing rate. Alternatively, θ_{z_i} is the marginal rate of substitution between z_i and money, or the implicit marginal

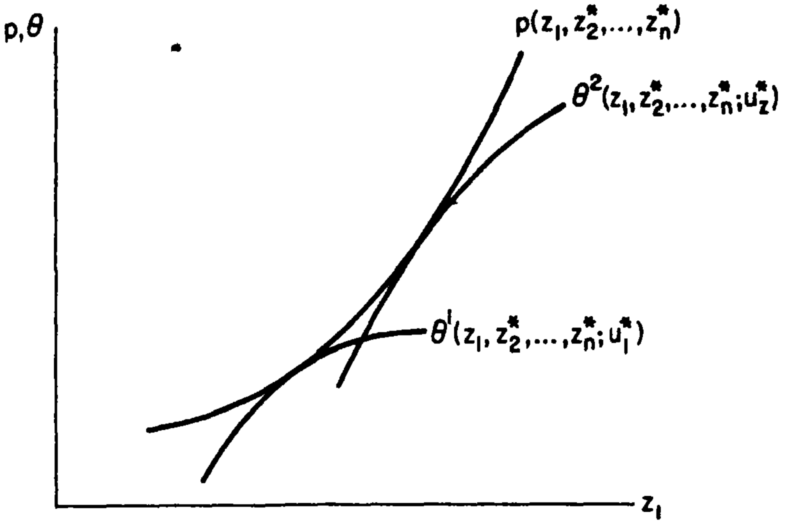


FIG 1

valuation the consumer places on z_i at a given utility index and income. It indicates his reservation demand price for an additional unit of z_i, which is decreasing in z_i

The amount the consumer is willing to pay for z at a fixed utility index and income is $\theta(z, u, y)$, while $p(z)$ is the minimum price he must pay in the market Therefore, utility is maximized when $\theta(z^*, u^*, y) = p(z^*)$ and $\theta_{z_i}(z^*, u^*, y) = p_i(z^*), i = 1, \ldots, n$, where z^* and u^* are optimum quantities. In other words, optimum location on the z-plane occurs where the two surfaces $p(z)$ and $\theta(z, u^*, y)$ are tangent to each other. One dimension of consumer equilibrium is illustrated in figure 1, where the surfaces have been projected onto the $\theta - z_1$ plane cut at $(z_2^*, \ldots, z_n^*)$ A family of indifference curves, of which only one member (at u^*) is shown, is defined by $\theta(z_1, z_2^*, \ldots, z_n^*, u, y)$. Two different buyers are shown in the figure, one with value function θ^1 and the other with θ^2 The latter purchases a brand offering more z_1 [3]

In general, far less can be said than in the standard analysis about comparative statics, because the budget constraint is nonlinear Differentiate θ_{z_i} with respect to u, $\theta_{z_iu} = (U_xU_{xz_i} - U_{z_i}U_{xx})/U_x^2$, the numerator of which is recognized as determining the sign of the income elasticity of demand for "good" z_i in standard theory when the other components of z are "held constant." If all these derivatives are positive (z_i is "normal" in this restricted sense for all i), the gradient of θ unambiguously

---

[3] Lewis (1969) employs a similar construction in analyzing the problem of hours of work as a tied sale Jobs offer a fixed wage-hour package, which varies from job to job. The market establishes a function relating wages and hours on which both workers and employers base their decisions.

increases as u increases Additional income always increases maximum attainable utility Hence if $p(z)$ is convex and sufficiently regular everywhere, we might expect higher income consumers to purchase greater amounts of all characteristics Only in that case would it be true that larger income leads to an unambiguous increase in the overall "quality" consumed, and differentiated products' markets would tend to be stratified by income However, in general there is no compelling reason why overall quality should always increase with income. Some components may increase and others decrease (cf Lipsey and Rosenbluth 1971) Be that as it may, a clear consequence of the model is that there are natural tendencies toward market segmentation, in the sense that consumers with similar value functions purchase products with similar specifications This is a well-known result of spatial equilibrium models. In fact, the above specification is very similar in spirit to Tiebout's (1956) analysis of the implicit market for neighborhoods, local public goods being the "characteristics" in this case He obtained the result that neighborhoods tend to be segmented by distinct income and taste groups (also, see Ellickson 1971) That result holds true for other differentiated products too

Allowing a parameterization of tastes across consumers, the utility function may be written $U(x_1, z_1, \quad, z_n; \alpha)$, where α is a parameter that differs from person to person Equilibrium value functions depend on both y and α A joint distribution function $F(y, \alpha)$ is given in the population at large, and equilibrium of all consumers is characterized by a family of value functions whose envelope is the market hedonic or implicit price function

The model is easily expanded to include several quantities, so long as consumers are restricted to purchasing only one model Following Houthakker (1952), the utility function becomes $U(x_1, z_1, \quad, z_n, m)$, where m is the number of units consumed of a model with characteristics z The constraint is $y = x + mp(z)$, and necessary conditions become

$$\frac{\partial U}{\partial m} = -p(z)U_x + U_m = 0, \tag{4}$$

$$\frac{\partial U}{\partial z_i} = -mp_i(z)U_x + U_{z_i} = 0 \tag{5}$$

The value function is still defined as the amount a consumer is willing to pay for z at a fixed utility index but now with the proviso that m is optimally chosen That is, $\theta(z_1, \quad, z_n)$ is defined by eliminating m from

$$u = U(y - m\theta, z_1, \quad, z_n, m)$$

$$U_m/U_x = \theta$$

Again, θ_{z_i} is proportional to U_{z_i}/U_x The logic underlying figure 1 remains intact, and it can just as well serve for this case. However, second-order conditions are now more complex. For example, convexity of $p(z)$ is no longer sufficient for a maximum as it was in the case where m was restricted to be unity. Also, it is necessary to employ stronger assumptions than those used above if the value function θ is to be concave

Note there is no question of monopsony involved here. Consumers act competitively in spite of the fact that marginal cost of quality, $p_i(z)$, is not necessarily constant—it is increasing in figure 1—because as many units as desired of any brand can be purchased without affecting prices The function $p(z)$ is the same for all buyers and independent of m

B *The Production Decision*

Having set up the formal apparatus above, we give a symmetrical and consequently brief account of producers' locational decisions. What package of characteristics is to be assembled? Let $M(z)$ denote the number of units produced by a firm of designs offering specification z. The discussion is limited to the case of nonjoint production, in which each production establishment within the firm specializes in one design, and there are no cost spillovers from plant to plant Thus a "firm" is an arbitrary collection of atomistic production establishments, each one acting independently of the others Analytical difficulties arising from true joint production are noted in passing

Total costs in an establishment are $C(M, z; \beta)$, derived from minimizing factor costs subject to a joint production function constraint relating M, z, and factors of production The shift parameter β reflects underlying variables in the cost minimization problem, namely, factor prices and production function parameters. Assume C is convex with $C(0, z) = 0$ and C_M and $C_{z_i} > 0$ There are no production indivisibilities, and marginal costs of producing more units of a model of given design are positive and increasing Similarly, marginal costs of increasing each component of the design are also positive and nondecreasing (Ordinarily, there will be some technological constraints that limit the set of feasible locations on the plane) Each plant maximizes profit $\pi = Mp(z) - C(M, z_1, \ldots, z_n)$ by choosing M and z optimally, where unit revenue on design z is given by the implicit price function for characteristics, $p(z)$ [4]

---

[4] Our inability to treat joint production nontrivially yet simply stems from the spectrum-of-commodities assumption If a finite number (say v) of packages is available, it would be straightforward formally to specify a standard v-component multiple product cost function for the firm, and proceed on that basis In the present case, firms engage in joint production only insofar as they own establishments specializing in different packages. However, genuine joint production requires cost dependencies between production units within the firm: the firm must choose a function $M(z)$ describing an entire "product line" offered in the market. The entire function $M(z)$ is an argument in each plant's costs and total costs in turn are the sum (or integral) over all production establishment costs. A complete treatment requires use of functional analysis and is beyond the scope of this paper

Accession numbers

Again, firms are competitors and not monopolists even though marginal costs of attributes $p_i(z)$ are not necessarily constant because all establishments observe the same prices and cannot affect them by their individual production decisions. $p(z)$ is independent of M.

Optimal choice of M and z requires

$$p_i(z) = C_{z_i}(M, z_1, \ldots, z_n)/M, \qquad i = 1, \ldots, n \tag{6}$$

$$p(z) = C_M(M, z_1, \ldots, z_n) \tag{7}$$

At the optimum design, marginal revenue from additional attributes equals their marginal cost of production per unit sold. Furthermore, quantities are produced up to the point where unit revenue $p(z)$ equals marginal production cost, evaluated at the optimum bundle of characteristics. As above, convexity of C does not assure second-order conditions due to nonlinearity of $p(z)$, and some stronger conditions, assumed to be satisfied in what follows, are required (see Intriligator 1971).

Symmetrically with the treatment of demand, define an *offer* function $\phi(z_1, \ldots, z_n, \pi, \beta)$ indicating unit prices (per model) the firm is willing to accept on various designs at constant profit when quantities produced of each model are optimally chosen. A family of production "indifference" surfaces is defined by ϕ. Then $\phi(z_1, \ldots, z_n; \pi, \beta)$ is found by eliminating M from

$$\pi = M\phi - C(M, z_1, \ldots, z_n) \tag{8}$$

and

$$C_M(M, z_1, \ldots, z_n) = \phi, \tag{9}$$

and solving for ϕ in terms of z, π, and β. Differentiate (8) and (9) to obtain $\phi_{z_i} = C_{z_i}/M > 0$ and $\phi_\pi = 1/M > 0$.

The marginal reservation supply price for attribute i at constant profit, assumed increasing in z_i, is ϕ_{z_i}. Again convexity of C does not always guarantee $\phi_{z_i z_i} > 0$. Since ϕ is the offer price the seller is willing to accept on design z at profit level π, while $p(z)$ is the maximum price obtainable for those models in the market, profit is maximized by an equivalent maximization of the offer price subject to the constraint $p = \phi$. Thus maximum profit and optimum design satisfy $p_i(z^*) = \phi_{z_i}(z_1^*, \ldots, z_n^*, \pi^*, \beta)$, for $i = 1, \ldots, n$, and $p(z^*) = \phi(z_1^*, \ldots, z_n^*, \pi^*, \beta)$. Producer equilibrium is characterized by tangency between a profit-characteristics indifference surface and the market characteristics–implicit price surface.

One dimension of the solution is depicted in figure 2, where

$$\phi(z_1, z_2^*, \ldots, z_n^*; \pi, \beta)$$

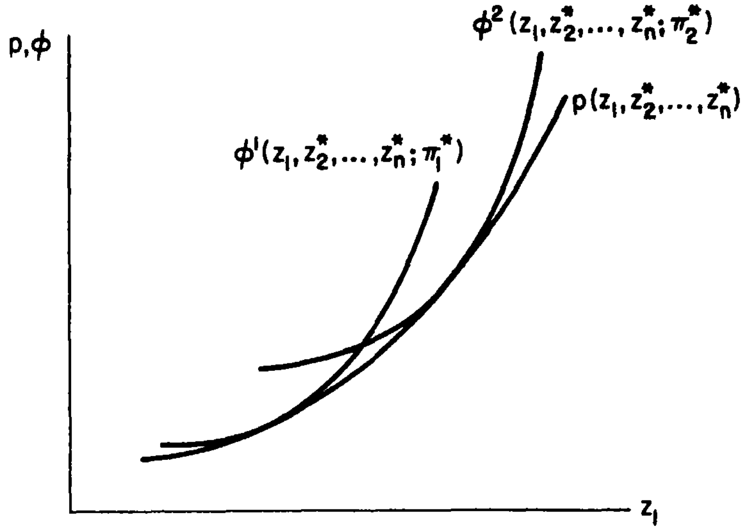


Fig 2

defines a family of curves on the $z_1 - \phi$ plane cut through the indifference surface at the optimum values of the other attributes Only one member is shown in the figure The curve labeled ϕ^1 refers to a production unit possessing production and cost conditions making it well suited to produce lesser amounts of z_1, while the one labeled ϕ^2 refers to a firm with a comparative advantage at producing higher values of z_1 That is, the two plants have distinct values of the parameter β More generally, there is a distribution of β across all potential sellers Let $G(\beta)$ represent that distribution. Then producer equilibrium is characterized by a family of offer functions that envelop the market hedonic price functions

What is the empirical content of β? It is anything that shifts cost conditions among firms Thus, differences in factor prices are one possibility For example, many products are produced in several countries and are traded on national markets (for examples, see Griliches [1971], chap 5) There is no reason to assume equalization of factor prices in these cases. More generally, anything allowing identification of conventional multiproduct production functions in cross-section data serves to provoke differences in β Factor price differences across states or regions within a country often serve this purpose and do so here as well Second, differences in "technology," as reflected by typically unmeasured, firm-specific factors of production, also act as supply shifters across firms. For example, agricultural production function research often treats education of the farm operator in this manner Firm-specific R&D expenditure as well as the phenomena of progress-function-learning also serve these purposes.

C What Do Hedonic Prices Mean?

An answer to the question is an immediate application of the above analysis Superimpose figure 2 onto figure 1 In equilibrium, a buyer and seller are perfectly matched when their respective value and offer functions "kiss" each other, with the common gradient at that point given by the gradient of the market clearing implicit price function $p(z)$ Therefore, observations $p(z)$ represent a joint envelope of a family of value functions and another family of offer functions An envelope function by itself reveals nothing about the underlying members that generate it; and they in turn constitute the generating structure of the observations. Some qualifications are necessary however (*a*) Suppose there is no variance in β and all firms are identical Then the family of offer functions degenerates to a single surface, and $p(z)$ must be everywhere identical with a unique offer function Price differences between various packages are exactly equalizing among sellers because offer functions are constructed at constant profit A variety of packages appear on products markets to satisfy differences in preferences among consumers, and the situation persists because no firm finds it advantageous to alter the quality content of its products (*b*) Suppose sellers differ, but buyers are identical Then the family of value functions collapses to a single function and is identical with the hedonic price function Observed price differences are exactly equalizing across buyers, and $p(z)$ identifies the structure of demand.

III. Existence of Market Equilibrium

Analysis of consumer and producer decisions has proceeded on the assumption of market equilibrium This section demonstrates some details of equilibrium price and quantity determination Market quantity demanded for products with characteristics z is $Q^d(z)$, and $Q^s(z)$ is market quantity supplied with those attributes It is necessary to find a function $p(z)$ such that $Q^d(z) = Q^s(z)$ for all z, when buyers and sellers act in the manner described above The fundamental difficulty posed by this problem is that $Q^d(z)$ and $Q^s(z)$ depend on the entire function $p(z)$ For example, suppose quantities demanded and supplied at a particular location do not match at prevailing prices The effect of a change in price at that point is not confined to models with those particular characteristics but induces substitutions and locational changes everywhere on the plane. A very general treatment of the problem is found in Court (1941), and our discussion is devoted to some examples. These examples have been chosen for their simplicity but illuminate the problem and illustrate most of the basic issues. In contrast to the rest of the paper, discussion is specialized to the case where goods are described by exactly one attribute (i.e., $n = 1$) Therefore z_1 represents an unambiguous measure of "quality." When $n = 1$, the location surface degenerates to a line rather than a plane, and products are unequivocally ranked by their z content.

A Short-Run Equilibrium

Consider a short-run equilibrium in which firms have geared up for the quality (z_1) of goods they can produce and are only capable of varying quantities. The horizon is sufficiently short so that new entry is precluded, and the distribution of firms by quality is given as an initial condition. The market reveals an implicit price function $p(z_1)$, and each firm determines the quantity it supplies to the market according to condition (7). Market supply in a small interval dz_1 near quality z_1 is found by weighting firm supply by the quality distribution function. Consumers differ in tastes and income, but all determine optimal quality and quantity as in (4) and (5). Market demand near any quality z_1 is found by using the conditions of consumer equilibrium to transform the distribution of tastes and income into a distribution of qualities demanded and weighting individual quantities demanded by the resulting distribution of qualities. Finally, setting demand equal to supply yields a differential equation in p and z_1 that must be satisfied by market equilibrium, subject to some boundary conditions.

To be specific, assume that $C(N, z) = (a/2)M^2z_1^2$ for all firms. Also, suppose firms are uniformly distributed by the characteristic z_1: $g(z_1)dz_1 = kdz_1$ for $z_{1s} \leq z_1 \leq z_{1l}$, where k is a constant and z_{1l} and z_{1s} are exogenously determined upper and lower limits of the product line. Apply equation (7) to obtain firm supply $M(z_1) = p/az_1^2$, since qualities cannot be varied by assumption. Therefore,

$$Q^s(z_1)dz_1 = g(z_1)M(z_1)dz_1 = [(k/a)p(z_1)/z_1^2]dz_1 \tag{10}$$

Assume a fixed number of consumers in the population and that only one unit per customer of the optimal model is purchased. Consumers have the same income, and utility is linear in x and z_1, with the marginal rate of substitution, ρ, varying from person to person. Maximize $U(x, z_1) = x + \rho z_1$ subject to $y = x + p(z_1)$. Each consumer purchases a brand for which $dp/dz_1 = p'(z_1) = \rho$. In this case the value functions of figure 1 are straight lines with a different slope, ρ, for each person. The marginal condition characterizes consumer choice so long as $p'' > 0$, which will be shown to be true. Suppose ρ is distributed uniformly, $f(\rho)d\rho = bd\rho$ for $\rho_s \leq \rho \leq \rho_l$, where b is a constant and ρ_l and ρ_s are, respectively, the largest and smallest marginal rates of substitution in the population. Use the marginal condition $p' = \rho$ to transform $f(\rho)d\rho$ into

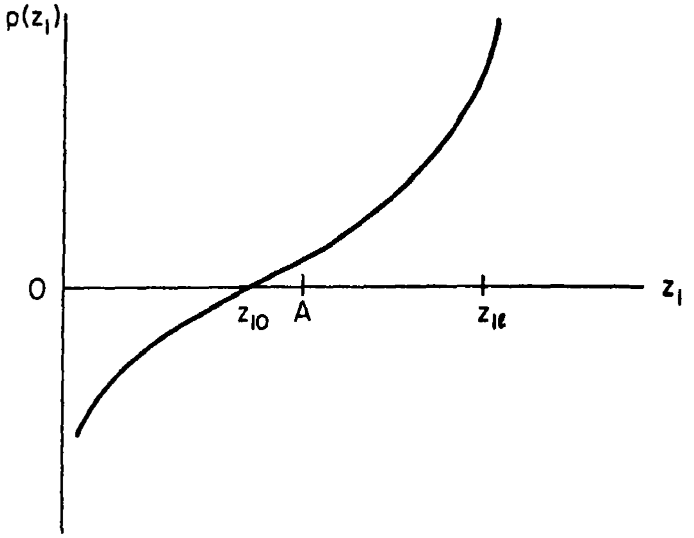


FIG 3

a distribution of z_1 Then

$$Q^d(z)dz = f(z_1)\left|\frac{dp}{dz_1}\right| dz_1 = bp''(z_1)dz_1 \tag{11}$$

Price must clear the market at every quality Equating (10) and (11), $p(z_1)$ must satisfy the differential equation

$$(k/ba)p/z_1^2 = d^2p/dz_1^2 \tag{12}$$

Equation (12) is a special case of what is called "Euler's equation" and has a known solution of the form

$$p = c_1 z_1^r + c_2 z_1^s, \tag{13}$$

where c_1 and c_2 are constants determined by the boundary conditions and r and s are defined by $r^2 - r - (a/bk) = 0$ $r = (1 + \sqrt{1 + 4a/bk})/2$ and $s = (1 - \sqrt{1 + 4a/bk})/2$ The parameters r and s are real numbers and $r > 0$ and $s < 0$ Furthermore, $p'(z_1)$ would not be positive throughout its range unless $c_1 > 0$ and $c_2 < 0$, and consumers could not be interior at those points Equation (13) is graphed in figure 3 on that assumption. Note that p in (13) exhibits an inflection point at $z_{10} = (-c_1/c_2)^{1/(r-s)}$, and it so happens that $p(z_{10}) \equiv 0$. Therefore $p'' > 0$ for $z_1 > z_{10}$.

Boundary conditions.—Competition requires there be no masses of consumers at any quality, for there are few sellers located at any point and they would otherwise add nonzero weight to the market. As seen in Section II, consumers with high values of ρ buy higher-quality models, and it must be true that those for whom $\rho = \rho_l$ purchase the highest quality available Otherwise prices of quality z_{1l} would fall, a great mass of consumers would switch over to them, driving the price back up and causing those buyers to relocate again Therefore, one boundary condition is $p'(z_{1l}) = \rho_l$, or

$$\rho_l = rc_1 z_{1l}^{r-1} + sc_2 z_{1l}^{s-1} \tag{14}$$

The other boundary condition is found by examining the lower end of the line The following three cases cover all relevant possibilities

1 $z_{1s} = 0$ and $\rho_s > 0$ Firms choose not to sell at negative prices (see fig 3) and all plants geared to produce qualities less than z_{10} (to be determined) shut down On the other hand, all consumers value z_1 at least as much as its minimum supply price (i e , zero) and it must be true that they all buy some value of z_1 Individuals for whom $\rho = \rho_s$ consume the lowest qualities appearing on the market, for if they chose qualities greater than z_{10}, prices of models in the neighborhood of z_{10} would fall to zero, inducing low ρ customers to relocate there and driving their prices back up. Thus a second boundary condition is $p'(z_{10}) = \rho_s$, or

$$\rho_s = rc_1 z_{10}^{r-1} + sc_2 z_{10}^{s-1} \tag{15}$$

The parameters z_{10}, c_1, and c_2 are determined by equations (14) and (15) plus the definition of z_{10} It can be shown that $c_1 > 0$ and $c_2 < 0$, as required by the second-order conditions of consumer equilibrium Therefore, the equilibrium hedonic price function appears as a portion of the curve in figure 3 in the interval (z_{10}, z_{1l})

2 If $\rho_s = 0 = z_{1s}$, all producers must be in the market, and it follows that $z_{10} = 0$ This only is possible if $p'(0) = \rho_s = 0$ and c_2 must be zero In this case price is a log-linear function of quality

3 $z_{1s} > 0$ and $\rho_s = 0$ Now some consumers do not value z_1 very highly, and there is a definite limit to the smallest amount available Clearly, $p(z_{1s})$ must exceed zero and some consumers must be driven out of the market, finding it optimal not to consume the product at all If not, consumers with small values of ρ would mass on z_{1s} (there would be a corner solution there), adding finite weight to the market and causing $p(z_{1s})$ to explode Using the budget constraint, the market rate of exchange between not buying at all and buying z_{1s} is $[y - p(z_{1s})]/z_{1s}$ and must equal the slope of the value function for buyers at that (extensive) margin. That is, the condition $[y - p(z_{1s})]/z_{1s} = p'(z_{1s})$ replaces equation (15)—after substituting for p and p' from (13)—in the determination of c_1 and c_2 The hedonic price function also can be illustrated in figure 3 as the portion of the curve between the points such as those marked $A(= z_{1s})$

and z_{1i}. Again, c_1 and c_2 have the correct signs and the second-order conditions are fulfilled.

A second type of short-run equilibrium could be considered in which existing firms can alter qualities as well as quantities of their products. When there is a distribution of cost functions, it is necessary to proceed analogously to the treatment of demand in the example above. For example, costs might be described by $(a/2)N^2z_1^\lambda$ with λ varying across firms. Then $(\lambda/2) = z_1p'/p$ is used to transform the distribution of λ into a distribution of qualities supplied. The resulting distribution weights firm quantities supplied in the determination of market supply at any quality. A little experimentation will show that the differential equation resulting from setting $Q^d(z_1) = Q^s(z_1)$ is nonlinear in most cases, and closed solutions are not always feasible.

B. Long-Run Equilibrium

Firms may vary qualities at will and also construct establishments of optimum size. No entry restrictions imply the absence of profit ($\pi^* = 0$) and long-run offer price for each firm must satisfy $\phi(z, \beta) = C(M, z, \beta)/M$. Plants are constructed to produce models of quality z at minimum cost. Hence scale economies are exhausted under competition and the optimum production unit occurs where $C(M, z, \beta)$ is linear in M, variations of quantity being achieved by changes in the number of establishments. Let $h(z, \beta)$ represent minimum average cost of z for an establishment of optimum size. Then $C(M, z, \beta) = Mh(z, \beta)$ in the long run. Therefore $\phi = h(z, \beta)$ and $p(z) = h(z, \beta)$ is the equilibrium condition for maximum profit and $p(z)$ is completely determined by supply, or by the envelope of the family $h(z, \beta)$ with respect to β. Generalization to n characteristics is obvious in this case.

IV. An Identification Problem

Section III demonstrated that complete solutions for $p(z)$ and the distribution of qualities traded sometimes can be obtained if sufficient a priori structure is imposed on the problem. However, it is not always possible to proceed in that manner. In general, the differential equation defining $p(z)$ is nonlinear and it may not be possible to find closed solutions. Moreover, a great deal of structure must be imposed. For example, the distribution of income follows no simple law throughout its range, making it difficult to specify the problem completely. Finally, partial-differential equations must be solved when there is more than one characteristic. This section sketches an alternative and more efficient procedure, based on the analysis of Section II.

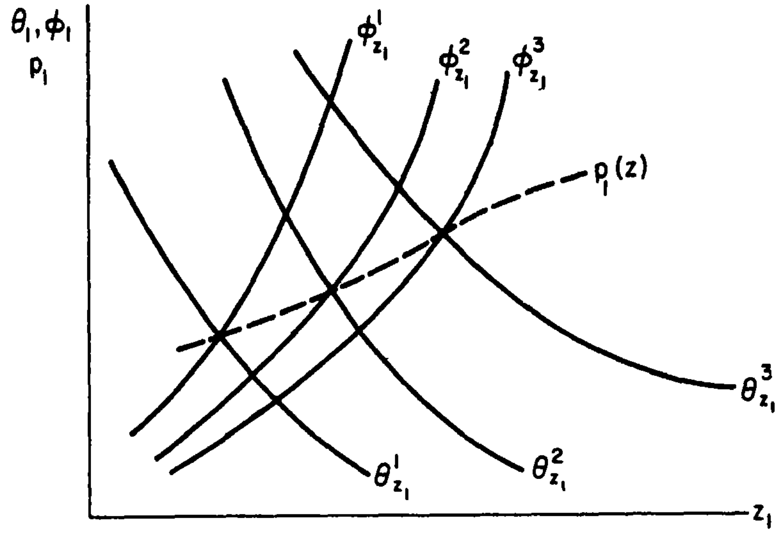


FIG 4

As shown above, derivatives of a consumer's value function, θ_{z_i}, are proportional to marginal rates of substitution. They are reservation-demand prices for additional amounts of z_i at a constant utility index. Therefore $\{\theta_{z_i}(z)\}$ are the inverses of a set of ordinary compensated demand functions for the z_i's. The marginal cost of z_i to the consumer is $p_i(z)$, and optimal z is determined where marginal costs equal marginal values. One dimension of these marginal concepts is illustrated in figure 4. The curves labeled $\theta^j_{z_i}$ are derivatives of θ^j in figure 1 and reflect compensated demand functions for various buyers. The dashed line labeled $p_1(z)$ is the common marginal cost confronting all buyers. Consumer choice is given by the intersection of demand and marginal cost. It should be emphasized that the functions $\theta_{z_i}(z)$ are compensated demand prices (real income held constant) and can only be derived once equilibrium is determined, as in Section II. For example, a new equilibrium resulting from an exogenous shift in p would not always be given by the intersection of the new marginal costs, $p_1(z)$, and the initial compensated demand price functions. An exception occurs when $\theta_{z_i u} = 0$ and the family of surfaces $\theta(z, u)$, such as depicted in figure 1, are all parallel to each other: $\theta_{z_i u} = 0$ is equivalent to constant marginal utility of money and θ_{z_i} is unique and independent of u only in that case. If $\theta_{z_i u} \neq 0$, the shape and location of the $\theta^j_{z_i}$ functions are determined by the equilibrium conditions of Section II: tangency between $p(z)$ and $\theta^j(z, u^*)$.

A similar procedure applies to firms: θ_{z_i} is the reservation supply price of incremental z_i and reflects a profit-compensated supply function for characteristic z_i, p_i is the marginal revenue function for z_i facing each firm. One dimension of producer equilibrium is shown in figure 4 as the intersection of a set of compensated supply curves for various firms, θ^j_z, with a common marginal revenue function, $p_1(z)$.

Figure 4 reiterates the major conclusion of Section II in terms of derivatives of $p(z)$. Equilibrium is described by the intersection of supply and demand functions. However, income effects have been removed, in distinction to the typical case. Observed marginal hedonic prices merely connect equilibrium reservation prices and characteristics and reveal little about underlying supply and demand functions.

However, figure 4 suggests a method that can be used for estimation. In principle, data are available on designs purchased by buyers and also on their incomes and taste variables such as age, education, etc. Denote these empirical counterparts of α by a vector Y_1. Data are also potentially available on the characteristics' content of models produced by sellers and factor price and specific technological differences among them. Denote the empirical counterparts of β by a vector Y_2. Following figure 4, let $F_i(z, Y_1)$ represent the marginal demand price for z_i and $G_i(z, Y_2)$ represent the marginal supply price. Ignoring random terms, the model to be estimated can be written as

$$p_i(z) = F^i(z_1, \ldots, z_n, Y_1) \quad \text{(demand)}, \tag{16}$$

$$p_i(z) = G^i(z_1, \ldots, z_n, Y_2) \quad \text{(supply)}, \tag{17}$$

for $i = 1, \ldots, n$, where p_i and z_i are all jointly dependent variables and Y_1 and Y_2 are exogenous demand and supply shift variables. The $2n$ equations determine the $2n$ endogenous variables p_i and z_i. Estimation requires a two-step procedure. First, estimate $p(z)$ by the usual hedonic method, without regard to Y_1 and Y_2. That is, regress observed differentiated products' prices, p, on all of their characteristics, z, using the best fitting functional form. This econometrically duplicates the information acquired by agents in the market, on the basis of which they make their decisions. Denote the resulting estimate of the function $p(z)$ by $\hat{p}(z)$. Next, compute a set of implicit *marginal* prices, $\partial p(z)/\partial z_i = \hat{p}_i(z)$ for each buyer and seller, evaluated at the amounts of characteristics (numerical values of z) actually bought or sold, as the case may be. Finally, use estimated marginal prices $\hat{p}_i(z)$ as endogenous variables in the second-stage simultaneous estimation of equations (16) and (17). Estimation of marginal prices plays the same role here as do direct observations on prices in the standard theory and converts the second-stage estimation into a garden variety identification problem. There are four cases to consider:

1. There is no variance in β and cost conditions are identical across firms. The variables Y_2 drop out of equation (17) and $\hat{p}(z)$ identifies the

offer function Similarly, the sample observations on $p_j(z)$ and the z_i identify compensated supply functions Suppose several cross sections for different years are available and firms' production functions have been subject to technical change Then within-year hedonic price regressions identify supply conditions for each year Changes in marginal prices and qualities induced by changing technology and cost conditions between years approximately sweep out the structure of preferences and compensated demand functions (with due qualification for the nonconstancy of the marginal utility of money)

2 If buyers are identical, but sellers differ, Y_1 drops out of (16) and single cross-sectional observations trace out compensated demand functions

3 If buyers are identical and so are sellers, offer and value functions are tangent at a single point, and only one quality appears on the market The observations degenerate to a single point, there is no product differentiation and no problem

4 In general there is both a distribution of buyers and another distribution of sellers Both Y_1 and Y_2 have nonzero variance, and the usual identifying rank and order conditions apply A necessary prior condition for estimation is that $p(z)$ be nonlinear at stage one For if $p(z)$ happens to be linear, $p_i(z)$ are constants, independent of qualities traded, and display zero variance across sample observations As shown above, linearity of $p(z)$ is unlikely so long as there is increasing marginal cost of attributes for sellers and it is not possible to untie packages But it is obvious that the model does not apply if very few distinct products are actually traded

V. Price Indexes, Economic Welfare, and Legislated Restrictions

This section uses the model to analyze the welfare consequences of quality-standards legislation, a problem not easily handled by conventional methods The discussion clarifies issues in recent controversies regarding treatment of legislated standards in the construction of price indexes For example, how should mandatory installation of seat belts and air bags affect the automobile price index? For expository convenience, discussion is confined to the case of one attribute Generalization to several characteristics is immediate

A minimum quality standard means that $z \geq \bar{z}$, and brands containing less than $\bar{z}$ are prohibited from the market Assume constant returns to quantities (as in Section III B) Then the law is irrelevant for all consumers previously purchasing packages containing more than the legislated minimum The situation for a buyer whose choice is affected by the law is shown in figure 5: z^* was the original choice, whereas $\bar{z}$ is chosen after the law has been passed, since z^* is no longer available The

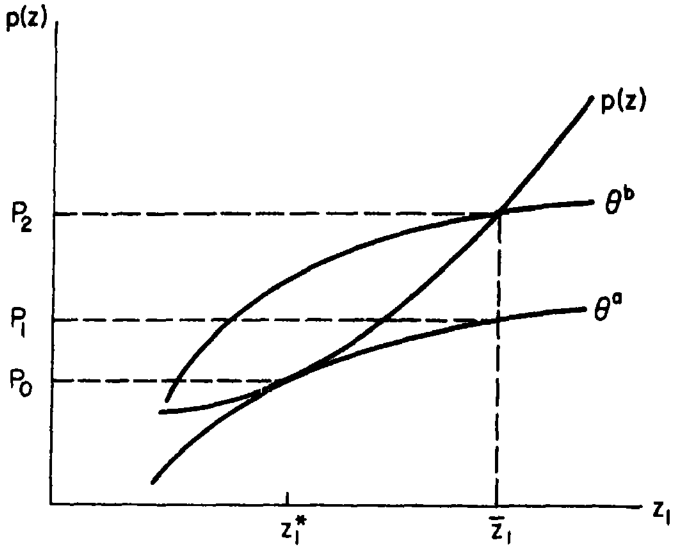


FIG 5

minimum attainable value function has shifted from θ^a to θ^b, and the consumer is worse off (see eq [2])

Choose the distance $\Delta P = P_2 - P_1$ as a monetary measure of the loss in welfare Since $\partial\theta/\partial y = 1$, ΔP is the bribe necessary for the consumer to purchase $\bar{z}$ when z^* was available Clearly, this measure is not unique (i.e , if compensation is evaluated at a different amount of z) unless $\theta_{z_1 u} = 0$ The welfare loss can be estimated from the implicit price and bid functions The distance $P_2 - P_0$ is given by

$$\int_{z_1^*}^{\bar{z}_1} p_1(z)dz,$$

or the area under marginal cost from z^* to $\bar{z}$, and is shown in figure 6 as $z_1^*ab\bar{z}_1$. It represents the social opportunity cost of additional resources necessary to produce $\bar{z}$ instead of z^* The integral

$$\int_{z^*}^{\bar{z}} \theta_1(z)dz,$$

or the area under a compensated demand function (compensated at the original level of real income) between z^* and $\bar{z}$ in figure 6 ($z_1^*ac\bar{z}_1$) measures the amount the consumer would have paid for the increment ($\bar{z}_1 - z_1^*$) at the unrestricted level of welfare. It measures $P_1 - P_0$ in figure 5 and represents the benefit of the restriction. The difference

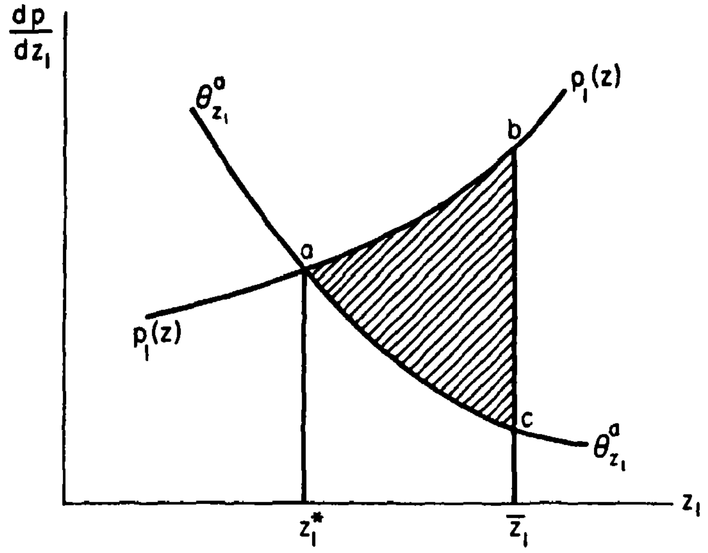


FIG 6

between costs and benefits is given by $P_2 - P_1$, or the difference between the areas under the marginal cost and compensated demand functions, the shaded area in figure 6 In the general case of several attributes, ΔP must be measured by a line integral Otherwise, everything else is unchanged

When the marginal utility of money is constant, ΔP is unique and the price restriction is equivalent to an additive increase in implicit prices in amount ΔP everywhere In figure 5, $\theta_{z_M} = 0$ means that all value functions are parallel, and if the budget constraint was $y = x + p(z) + \Delta P$ instead of $y = x + p(z)$, the consumer would have arrived exactly at θ^b of his own free choice The real price of the characteristic has risen because choices are restricted, and the price index should rise to reflect that fact A natural measure of the real price increase imposed by the law is a weighted average of terms such as ΔP (including buyers for whom $\Delta P = 0$), where the weights are expenditure shares among all consumers.[5] This measure overstates the loss insofar as the restriction actually forces some consumers completely out of the generic goods market since they escape the full loss ΔP Also, standard index number problems arise when the marginal utility of money is not constant

---

[5] A complete assessment of the law and its effect on the price index requires balancing the costs calculated above against any externality-induced social benefits of the restriction In our judgment, seat belts and air bags are in a different category than emission-control devices. In regard to the latter, the apparatus above can be used easily to analyze the effect of the European system of taxing engine displacement. An ad valorem tax increases average and marginal costs of packages with larger liter capacity, and the usual income and substitution effects apply: packages with smaller amounts of this and complementary characteristics (such as size of car) are purchased

VI. Conclusions

This paper has drawn out the observational consequences of the construct of implicit markets for characteristics embodied in differentiated products When goods can be treated as tied packages of characteristics, observed market prices are also comparable on those terms The economic content of the relationship between observed prices and observed characteristics becomes evident once price differences among goods are recognized as equalizing differences for the alternative packages they embody Here, as elsewhere, price differences generally are equalizing only on the margin and not on the average Hence, estimated hedonic price-characteristics functions typically identify neither demand nor supply In fact, those observations are described by a joint-envelope function and cannot by themselves identify the structure of consumer preferences and producer technologies that generate them

The formal analysis is complicated by the fact that budget constraints are nonlinear Consequently, it is not surprising that far weaker theorems than usual apply However, a feasible econometric procedure for estimating the underlying generating structure has been derived through the use of derivative transformations When constraints are nonlinear, marginal prices serve the same role as average prices do in the linear case Finally, the essential spatial context of the problem means that substitution and income effects must be more carefully distinguished than usual Indeed, here is a major practical instance where compensated demand and supply functions become the relevant fundamental concepts. These compensated functions are estimated by the econometric method and measures of consumer and producer surplus can be derived directly from them We anticipate that the basic conceptual framework outlined above will have a variety of applications to many practical problems involving equilibrium in cross-section data

The analysis has been simplified by assuming divisibility in production Generalization has to incorporate nonconvexities, and discontinuities must result When nonconvexities are not small relative to the market, it is obvious that only isolated locations on the characteristics surface will be filled In other words, such a generalization will naturally incorporate the case of monopolistic competition, and observed "distances" (in terms of characteristics) between differentiated products will be endogenously determined The methods employed above do not carry through because certain nonmarginal decisions must be analyzed, and far more sophisticated techniques are required

References

Alonzo, William. *Location and Land Use.* Cambridge, Mass.: Harvard Univ. Press, 1964.

Becker, Gary S "A Theory of the Allocation of Time." *Econ. J* 75 (September 1965). 493–517.

Court, Louis M "Entrepreneurial and Consumer Demand Theories for Commodity Spectra." *Econometrica* 9, no. 1 (April 1941) 135–62, no 2 (July–October 1941): 241–97.

Ellickson, Bryan. "Jurisdictional Fragmentation and Residential Choice" *A.E.R.* 61 (May 1971) 334–39

Gordon, Robert J "The Measurement of Durable Goods Prices "Mimeographed Nat. Bur Econ. Res., 1973

Griliches, Zvi, ed *Price Indexes and Quality Change* Cambridge, Mass. Harvard Univ. Press, 1971

Houthakker, H S "Compensated Changes in Quantities and Qualities Consumed" *Rev. Econ. Studies* 19, no 3 (1952)· 155–64

Intriligator, Michael D *Mathematical Optimization and Economic Theory* Englewood Cliffs, N J Prentice-Hall, 1971.

Lancaster, Kelvin J "A New Approach to Consumer Theory" *J P E.* 74 (April 1966) 132–56

Lewis, H Gregg "Interes del empleador en las horas de Trabajo del empleado" [Employer interests in employee hours of work] *Cuadernos de Economia*, Catholic Univ Chile, 1969

Lipsey, Richard G, and Rosenbluth, Gideon "A Contribution to the New Demand Theory A Rehabilitation of the Giffen Good" *Canadian J Econ* 4 (May 1971) 131–63

Muth, Richard F "Household Production and Consumer Demand Functions" *Econometrica* 34 (July 1966) 699–708

Ohta, Makoto, and Griliches, Zvi "Makes and Depreciation in the U S Passenger Car Market" Mimeographed. Harvard Univ, 1972

Tiebout, Charles M "A Pure Theory of Local Expenditure" *J P E.* 64 (October 1956) 416–24

A Diagrammatic Exposition of the Harberger Model with One Immobile Factor

Charles E. McLure, Jr.

Rice University

A four-quadrant diagram is used to exposit the Harberger model of tax incidence (Harberger 1962) in the case in which one of two factors is immobile between the two sectors Beyond its basically expository purpose, the paper offers two original contributions. First, the general equilibrium "supply curve" relating percentage changes in output to percentage changes in relative product prices is developed explicitly. The roles of various production parameters and factor immobility are thus highlighted Second, the diagrammatic exposition makes explicit one danger of using the Harberger model, written in terms of differentials, to examine tax changes of finite size.

I Introduction

In the past 10 years the incidence model presented by Arnold Harberger in his famous article in the June 1962 *Journal of Political Economy* has become the standard theoretical tool of economists concerned with the incidence of taxes in a general equilibrium setting [1] But many economists may find this analysis largely unintelligible because of the mathematical

---

This paper reports research related to AID contract csd-3302 to the Program of Development Studies at Rice University on the Distribution of Gains, Wealth, and Income from Development The author wishes to acknowledge the contributions Greg Ballentine, Ibrahim Eris, Paul Hanley, Wayne Thirsk, and an unusually meticulous referee for this *Journal* made through their careful reading and comments on an earlier draft As always, neither the Agency for International Development nor the individuals mentioned above should be held accountable for the views expressed here.

[1] See McLure (1972*a*) for a more detailed description and appraisal of the original Harberger model, its extensions, and applications

manipulations involved. Therefore a brief diagrammatic presentation of one version of the Harberger model is offered here in the hope of making that extremely useful model available to a wider audience In this paper only the case in which one factor is mobile between sectors and one factor immobile is presented, primarily because the original Harberger model with perfect mobility of both factors does not readily lend itself to the kind of diagrammatic treatment presented for the imperfect mobility case. Yet this exposition of the imperfect mobility case should assist in the understanding of even the complete mobility case Moreover, many policy problems involve time horizons in which analysts cannot reasonably assume that all factors are completely mobile, either between regions or industries This is especially true in the sphere of geographic incidence analysis In this paper both tax and expenditure incidence are examined In fact, it will simplify understanding of the model to consider expenditure incidence before tax incidence

In addition to the basically expository objective described above, this paper has two further purposes First, the "supply curve" relating percentage changes in the output of the taxed good to percentage changes in relative product prices is developed explicitly Use of this supply curve, in conjunction with the "demand curve," facilitates comparison with traditional partial equilibrium incidence analysis Moreover, we are able to see quite clearly how the various parameters of the model determine the supply response to changes in relative prices By comparison, none of the discussions of the Harberger model up until how has developed an explicit product supply relation Rather, they have focused primarily upon responses in markets for factors Of course, in the mathematical treatment of the model, all the equations are solved simultaneously The point is that the demand relation has never been juxtaposed against a supply relation derived from underlying conditions in factor markets and production functions [2]

Second, strictly speaking, the Harberger model is appropriate for the analysis of only small changes in taxes, since all the equations are written in terms of differentials. The diagrammatic exposition enables us to appreciate one of the pitfalls that threatens one who uses the Harberger model to analyze tax changes that are not small In particular, the model is developed initially under the assumption that the linearization of economic relationships inherent in the Harberger formulation in terms of differentials is acceptable. This is done solely as a matter of expository convenience But Section VI discusses the propriety of linearization when discrete changes in taxes are involved

---

[2] For a more detailed discussion of using the Harberger model to "go behind the supply curve" of traditional partial equilibrium analysis, see McLure (1972a)

The diagrams presented here are used primarily to explain the general structure and workings of the model and to isolate the crucial parameters of the analysis No attempt is made to present the results of the analysis in detail, except for polar cases, since anyone interested in the results for intermediate cases would undoubtedly turn to the algebraic formulation. Thus this paper may help readers to understand the more detailed algebraic version, but it is no substitute for it. Nevertheless, the link between the algebraic and diagrammatic expositions is made explicit by the textual development of the equations plotted in each diagram

Because this paper is intended to be primarily expository, the assumptions underlying the analysis are spelled out in detail But for the most part this is done as the various diagrams are developed, rather than as a group at the outset Thus the reader must be patient if he suspects that some crucial assumption is being neglected, hopefully it will come forth and be explained in due course Nevertheless, it will be convenient to describe several general features of the model at this point

First, the model is one in which two factors which are inelastically supplied in total (and by each individual) are employed in two competitive industries under constant returns to scale to produce two goods, x, y[3] Capital is assumed to be perfectly mobile between sectors Therefore, the price of capital services net of tax is the same in both sectors On the other hand, labor is assumed to be completely immobile between sectors, thus wage rates net of tax may differ between sectors

Second, in a model such as that employed here only changes in relative prices (as opposed to changes in absolute prices) can be determined in the absence of an explicit monetary equation But this in no way limits the usefulness of the model for incidence analysis, which is essentially a question of changes in relative factor returns (incidence on the sources side of income) and changes in relative product prices (incidence on the uses side of income) Thus we are free to choose one good or factor (or some weighted average) as the numeraire by setting the policy-induced change in its price equal to zero Though in general I prefer to use the untaxed product as the numeraire (in which case $dP_y = 0$ is the numeraire equation), the diagrammatic analysis is simplified by choosing capital services as our numeraire (so that the numeraire equation is $dP_K = 0$) Of course the outcome in terms of tax-induced changes in real income is in no way dependent upon the choice of numeraire equation Moreover, it will be convenient, but not essential, to define units of products and factors in such a way as to set all prices initially equal to unity. Thus any specific tax is also an ad valorem tax, and any change in price is also a fractional change.[4]

---

[3] Throughout this paper reference is to two sectors or industries In my earlier work I have used the model to investigate interregional incidence and related problems by dealing with two regions rather than two industries (see McLure 1969, 1970*a*, 1971*a*). This and other applications of the model are straightforward and hence are ignored in this paper.

Finally, the Harberger model is perhaps the comparative statics model par excellence in the sense that it can be solved for tax-induced changes from a preexisting (no-tax) equilibrium, but not for either the initial (no-tax) or the final (*cum*-tax) equilibrium solution. Thus in the diagrams of this paper all axes indicate percentage changes in variables, with the initial equilibrium corresponding to zero rates of change of all variables These changes are given initially in differential terms, following the original Harberger formulation Extension of the model to cases of discrete changes in variables is discussed in Section VI.

II. The Basic Diagram

The basic analytical tool developed here is the four-quadrant diagram presented in figure 1 It is derived on the assumption that there are no taxes and government spending in the economy

In quadrant *a* is plotted the "demand curve," D·

$$dQ_x/Q_x = -E(dP_x - dP_y) \tag{1}$$

(Ignore curve D_E for the moment, it is used in Section III) This description of demand differs from the standard demand curve in very special ways that must be made explicit First, the curve relates the percentage change in the demand for good x to the percentage change in the relative prices of goods x and y Thus it passes through the origin of the quadrant *a* diagram, and the reciprocal of the absolute value of its slope is the elasticity of demand for good x with respect to relative prices, E [5] Second, this elasticity of demand takes account of only substitution effects on the demand for good x resulting from changes in relative prices Income effects are ignored initially by using the very special assumption that at the margin all economic groups (including the government) spend their incomes in exactly the same way. Thus E is the market elasticity of demand for x with respect to relative prices, excluding first-order income effects

---

[4] For more detailed discussions of the issues raised in this paragraph, see Musgrave (1959, pp 217–21 and chap 15), Harberger (1962, pp. 219, 224), and McLure (1970*b*, 1971*b*, pp. 31–33)

[5] As drawn, the curve shows a constant demand elasticity, but it need not do so. This constancy of slope is not so bothersome in the specification of the demand relationship or even of the Σ_x curve of quadrant *c*, where a constant slope implies a constant elasticity of factor substitution. But for the curves in quadrants *b* and *d* and for the Σ_y curve in quadrant *c* the constant slope is in general quite unrealistic. This problem, which involves the analysis of differential changes, rather than discrete changes, in the Harberger model, is discussed further in Section VI below. For now the curves are all drawn as straight lines with full knowledge that, strictly speaking, this is unacceptable for discrete tax changes.

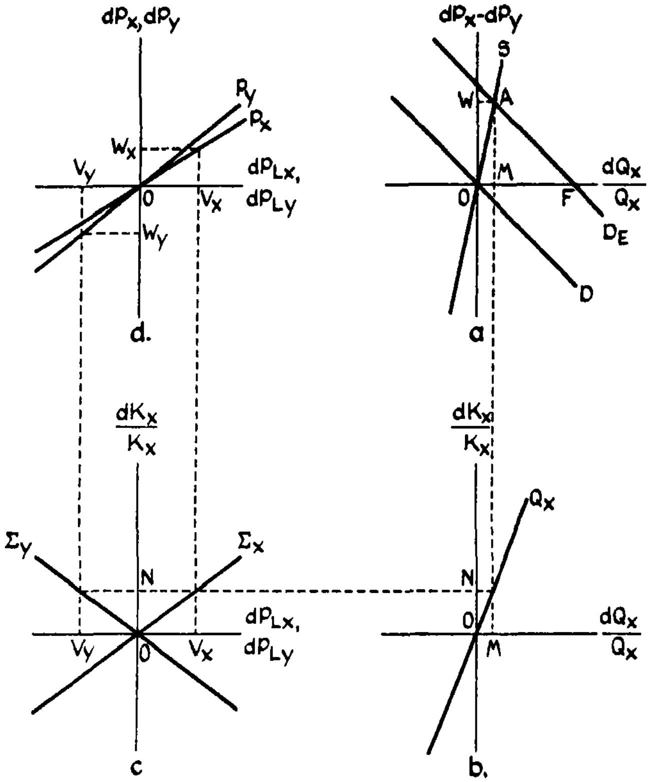


Fig. 1.—Derivation of supply curve and analysis of expenditure incidence

The second curve in quadrant *a*, *S*, which corresponds to the supply curve of traditional partial equilibrium analysis, is derived from the remaining quadrants. The remainder of this section is devoted to describing that derivation. The curve in quadrant *b* is described by the following equation:

$$dQ_x/Q_x = f_K\, dK_x/K_x + f_L\, dL_x/L_x = f_K\, dK_x/K_x \qquad (2)$$

It describes how output of good *x* changes if the amount of services of the mobile factor, capital, combined with the fixed amount of services of the immobile factor, labor, is changed. The curve Q_x relating dQ_x/Q_x, the percentage change in output of *x* to dK_x/K_x, the percentage change in capital services used in producing *x*, has a slope of $1/f_K$, where f_K can be interpreted alternatively as the elasticity of output with respect to this input in industry *x* or as capital's initial share of the output of industry *x*. Of course, this line, too, passes through the origin.

If production exhibits constant returns to scale, as is assumed in the Harberger model, factor proportions are uniquely related to the relative costs of the services of the two factors Stated differently, the possibility of changing factor proportions in production in response to changes in the relative costs of factor services is given by the elasticity of factor substitution in production In general, the possibilities of factor substitution in the two industries are given by the following two equations.

$$\frac{dK_x}{K_x} - \frac{dL_x}{L_x} = -\sigma_x(dP_K - dP_{Lx}) \tag{3}$$

and

$$\frac{dK_y}{K_y} - \frac{dL_y}{L_y} = -\sigma_y(dP_K - dP_{Ly}), \tag{4}$$

where σ_x and σ_y are the elasticities of factor substitution in the two industries Making use of the additional assumptions (*a*) that labor is immobile between sectors, (*b*) that the total supply of capital is fixed, and (*c*) that capital is chosen as the numeraire, so that $dP_K = 0$, we can rewrite equations (3) and (4) as

$$\frac{dK_x}{K_x} = \sigma_x\, dP_{Lx} \tag{3'}$$

and

$$\frac{dK_x}{K_x} = -\sigma_y \frac{K_y}{K_x}\, dP_{Ly} \tag{4'}$$

These equations are plotted in quadrant *c* as Σ_x and Σ_y As written, they show the changes in the utilization of capital services in industry *x* that would result from a given change in the cost of labor in each of the two industries, given that the price of capital services is held constant The slopes of these curves are σ_x and $-\sigma_y K_y/K_x$, respectively, and both curves pass through the origin

The fourth quadrant converts changes in the costs of factor services in the two industries into changes in product prices The basic equations are

$$dP_x = f_L\, dP_{Lx} + f_K\, dP_K \tag{5}$$

and

$$dP_y = g_L\, dP_{Ly} + g_K\, dP_K \tag{6}$$

In commonsense terms, these equations state that the percentage change in the price of a product is the weighted average of the percentage changes in the prices of the services of the two factors used to produce it The weights are the initial shares of output of the good in question accruing to the two factors. Since capital has been chosen as the numeraire, we can reduce these two equations to the following:

$$dP_x = f_L \, dP_{Lx} \tag{5'}$$

and

$$dP_y = g_L \, dP_{Ly} \tag{6'}$$

These are plotted in quadrant *d* as P_x and P_y As is true of all the curves in the initial equilibrium without taxes, these curves pass through the origin

Having derived the curves in quadrants *b–d* from the basic equations of the Harberger model, we can now use them to derive the supply curve for *x* with respect to relative prices The derivation is a straightforward exercise in the use of the diagrams First, we choose an arbitrary point, *M*, on the horizontal axis in quadrant *a* as a starting point Moving vertically to quadrant *b* we see that if output of good *x* is to increase by the fraction *OM*, the amount of capital services employed in industry *x* must increase by the fraction *ON* But this reallocation of capital implies a rise in the wage rate in industry *x* (given $dP_K = 0$) of OV_x and a fall in the wage rate in industry *y* of OV_y, as indicated by curves Σ_x and Σ_y in quadrant *c*

These changes in wage rates in the two industries are transmitted to product prices in the way described by curves P_x and P_y in quadrant *d* Thus the fractional increase in the wage rate of OV_x in industry *x* produces a fractional rise in the price of good *x* of OW_x Similarly, the price of good *y* falls by the fraction OW_y in response to the wage rate in the industry falling by the fraction OV_y Thus we have seen that the fractional increase in the output of good *x* of *OM* implies a change in the relative prices of goods *x* and *y* or W_yW_x [6] This change in relative prices can be plotted as *OW* in the quadrant *a* against *OM* to produce *A*, one point on the supply curve of good *x* with respect to relative product prices Repeating the procedures in this and the previous paragraph for other points on the horizontal axis in quadrant *a* produces the entire supply curve

Before we consider how taxes and expenditures affect the curves and values of the variables in the model, it will be convenient to digress briefly at this point to examine the determinants of the shape of the supply curve. First, we can describe the supply curve in general terms in the following way·

$$dQ_x/Q_x = S(dP_x - dP_y) \tag{7}$$

---

[6] Some readers may object that this derivation reverses the direction of causality in our usual way of thinking about supply curves, which is in terms of a given change in relative product prices giving rise to a certain change in output. The clockwise derivation described implies nothing about causality. It was chosen simply because it is substantially easier than a counterclockwise derivation, as the reader can easily ascertain for himself

Of course, S is the elasticity of supply of good x with respect to relative prices To obtain an expression for the elasticity of supply that utilizes information about conditions in factor markets and the production function we can repeat the reasoning of the last two paragraphs Thus

$$\begin{aligned}
dK_x/K_x &= dQ_x/Q_x \cdot f_K, \\
dP_{Lx} &= dK_x/K_x\sigma_x, \\
dP_{Ly} &= -K_x\, dK_x/K_x\sigma_y K_y, \\
dP_x &= f_L\, dP_{Lx}, \\
dP_y &= g_L\, dP_{Ly}, \\
dP_x - dP_y &= f_L\, dP_{Lx} - g_L\, dP_{Ly}, \\
&= f_L\, dK_x/K_x\sigma_x + g_L K_x\, dK_x/K_x\sigma_y K_y, \\
&= (f_l/\sigma_X + g_L K_x/K_y\sigma_y)\, dK_x/K_x, \\
&= (f_L/\sigma_x + g_L K_x/K_y\sigma_y)\, dQ_x/Q_x f_K
\end{aligned}$$

From the last expression we can calculate the elasticity of supply with respect to relative prices as

$$S = \frac{f_K\sigma_x\sigma_y}{f_L\sigma_y + g_L\sigma_x K_x/K_y} \tag{8}$$

Under the linearization assumption followed through Section V, S is a constant Thus curve S in quadrant a is a straight line

The determinants of the elasticity of supply can be seen directly from equation (8) to be (a) the elasticities of factor substitution in the two industries, (b) the allocation of the total capital stock between the two industries, and (c) labor's initial share of output in the two industries Partial differentiation of equation (8) reveals that *ceteris paribus* the supply of good x is more elastic the easier is factor substitution in either industry, the smaller is the fraction of the total capital stock devoted to production of good x, and the smaller is labor's share of output in each industry The same thing is revealed by inspection of figure 1 The steeper are curves Σ_x and Σ_y in quadrant c, the less is any change in the allocation of capital between the two industries reflected in relative factor prices and hence in relative product prices Thus the greater are σ_x, σ_y, and K_y/K_x, the greater is the elasticity of product supply with respect to relative product prices Similarly, *ceteris paribus*, the flatter are curve Q_x in quadrant b and curves P_x and P_y in quadrant d, the less effect a given change in the quantity of good x has upon relative product prices But flatness in any of these curves implies a low labor share in one industry

The economics behind these results is easy to appreciate First, if factor substitution in industry x is easy, capital can be absorbed with relatively little rise in the wage rate in that industry and hence little rise in the price of good x Similarly, if labor can easily be substituted for capital in industry y, the outflow of capital necessary to expand production of good x will not greatly lower the wage rate in industry y and the price of good y Essentially the same result occurs if relatively little of the total capital stock is used in industry x In such a case the capital inflow needed to expand production of good x can occur without greatly affecting factor proportions, relative factor returns, and the product price in industry y [7] Finally, the greater is capital's share in industry x, the smaller is the capital flow induced by a given fractional change in the output of that industry This, in turn, is reflected in a smaller change in wage rates and product prices in the two industries And the smaller is labor's share in either industry, the less responsive is the price of the product of that industry to a given change in the relevant wage rate

It may be worthwhile, to carry this reasoning further, to note the value of the supply elasticity for extreme values of the parameters just discussed If factor substitution is impossible in either industry, supply is completely inelastic Given the immobility of labor, output of x cannot expand unless factor substitution is possible in both industries

If in industry x factors are perfect substitutes, the elasticity of supply depends upon the elasticity of substitution in the other industry, factor shares in both industries, and the split of capital between the industries If perfect substitution characterizes production in industry y, the supply elasticity depends on the ease of factor substitution and relative factor shares in industry x Finally, if virtually all of the capital stock is used in industry x, the supply curve for that industry is quite inelastic But if almost none of the capital stock is used in the industry, the elasticity of supply depends crucially upon the elasticity of substitution and factor shares in the industry, just as when σ_y is infinite

Finally, it is instructive to calculate the elasticity of product supply for various values of the crucial parameters To simplify matters, one can calculate the elasticity assuming that virtually none of the capital stock is being utilized in industry x, since inspection of equation (8) reveals that this assumption yields the largest possible value of S, given the values of the other parameters Table 1 gives the values of S in the limit as K_x/K_y approaches zero, for the stated values of σ_x and factor shares in industry x One is immediately struck by the fact that these elasticities are all quite low, only at the extreme limits of the acceptable range of values for σ_x and factor shares does the elasticity of product supply equal or exceed unity.

---

[7] The numeraire assumption $dP_K = 0$ is not equivalent to an elastic supply of capital to industry x This elasticity, which depends upon the values of σ_y and K_x/K_y, is a matter of response to changes in relative factor returns and is independent of the choice of numeraire

TABLE 1

ELASTICITY OF PRODUCT SUPPLY WITH RESPECT TO RELATIVE PRICES, FOR VARIOUS VALUES OF σ_x, f_K, AND f_L (LIMIT AS $K_x/K_y \rightarrow 0$)

			VALUE OF σ_x						
f_K	f_L	f_K/f_L	0.2	0.4	0.6	0.8	1.0	1.2	1.5
0.1	0.9	0.11	0.02	0.04	0.07	0.09	0.11	0.13	0.17
0.2	0.8	0.25	0.05	0.10	0.15	0.20	0.25	0.30	0.38
0.3	0.7	0.43	0.09	0.17	0.26	0.34	0.43	0.51	0.64
0.4	0.6	0.67	0.13	0.27	0.40	0.53	0.67	0.80	1.00
0.5	0.5	1.0	0.20	0.40	0.60	0.80	1.00	1.20	1.50

Formula For $K_x/K_y = 0$, $S = f_K\sigma_x/f_L$

And for the most realistic assumptions of values of the parameters ($.6 \leq \sigma_x \leq 1.0$ and $.6 \leq f_L \leq .8$), the elasticity of supply falls within a range of .15–.67. All things considered, it seems likely that in industries and regions characterized by the kind of immobilities assumed here the elasticity of supply is probably no greater than $\frac{1}{3}$–$\frac{1}{2}$.[8] The elasticity of supply is, of course, lower than shown in table 1 if K_x/K_y is not near zero. For example, if $f_L = g_L$ and $\sigma_x = \sigma_y$, the elasticity of supply is only one-half as great as shown in table 1 if $K_x = K_y$, and it is only one-fourth as great if $K_x = 3K_y$.

In constructing the diagrams in this paper, what seem to be more or less reasonable values of the parameters are used. In particular, we use $\sigma_x = \sigma_y = .7$, $f_L = .6$, $g_L = .8$, and $K_x/K_y = 1$. (Different labor shares in the two industries are employed so that curves P_x and P_y in quadrant *d* do not coincide and K_x/K_y is not assigned a value near zero so that curve Σ_y in quadrant *c* will not be vertical.) This gives a value of S of .20.

---

[8] If we reverse the mobility assumptions to make capital the immobile factor and labor the mobile factor, as in the usual examples of the short run in price theory, the elasticity of supply is greater than shown in table 1 by a factor of $(f_L/f_K)^2$. But one can expect elasticities of factor substitution to be considerably higher ex ante or in long-run equilibrium (when capital is mobile) than they are once the capital stock is given. If both factors are mobile between sectors, the elasticity of supply of good *x* with respect to relative prices can be written as follows

$$S = \left[\sigma_y + \sigma_x\left(f_K \frac{L_x}{L_y} f_L \frac{K_x}{K_y}\right)\right] \Big/ \left[(f_K - g_K)\left(\frac{K_x}{K_y} - \frac{L_x}{L_y}\right)\right]$$

From this expression we can derive the usual results that in the perfect mobility case the supply response to a change in relative product prices can range from zero (if the two industries utilize the two factors in different proportions and factor substitution is completely impossible in both industries) to infinitely elastic (if the two factors are perfect substitutes in either industry or if the two industries utilize the two factors in the same proportions).

III. Expenditure Incidence

Thus far only the supply and demand curves and the derivation of the former have been given, no change from the initial equilibrium position has been analyzed In this and the next two sections the analytical framework developed in the last section is used for the comparative statics analysis of the distributional implications of tax, expenditure, and balanced-budget policies As noted earlier, it will be convenient to begin the investigation with the infrequently examined, but analytically simpler, question of the incidence of government expenditures Then we can proceed to the more complex questions of tax incidence and balanced-budget incidence in Sections IV and V

"Expenditure incidence" refers to the way government spending (as opposed to taxation) affects the distribution of income available for private use, by affecting relative factor and product prices [9] Government spending, financed in a neutral way, will have no effect on relative product prices, relative prices of factor services, and the distribution of income, unless at the margin the government spending pattern differs from the private pattern [10] But if the government devotes a large (smaller) share of its budget to good x than would have occurred at the margin in the private sector, the demand curve shifts to the right (left), and relative prices and the distribution of income change [11] We can trace this through the four-quadrant diagram

Consider an increase in government spending of dG In such a case the demand curve shifts by $(m_g - m_p)\, dG/Q_x$, where m_g and m_p are the public and private marginal propensities to consume good x (If m_g and m_p are equal, no shift occurs) Let us denote this shift by the shorthand notation dQ_G/Q_x The shift in the demand curve D is shown in figure 1, D_E being the shifted curve The intersection of the supply curve S and the shifted demand curve yields a new equilibrium position A, in which output of good x has increased by the fraction OM and the price of x has risen relative to that of y by OW[12]

---

[9] This concept of expenditure incidence is based upon the discussion in Musgrave (1959, pp 213–15) and is discussed more fully in McLure (1972*b*) It might be noted first that expenditure incidence, defined in terms of the effects of government spending on relative prices and real incomes, is closely akin to the concept of pecuniary externalities familiar to cost-benefit analysts, and second that the term "expenditure incidence" does not refer to the question of who benefits from government services That question is not generally amenable to the type of analysis presented here In this paper only the incidence of government purchases of final products is considered. The much more complicated problems of the incidence of direct governmental employment of factors, which is considered in McLure (1972*b*), is relatively difficult to fit into the diagrammatic analysis, though it is not impossible

[10] Reference to neutral financing can be interpreted to mean financing through a general sales or income tax (which in this model are equivalent) That such a tax is distributionally neutral is shown in the next section For a more complete discussion of distributionally neutral financing and expenditure patterns, see McLure (1971*b*, pp. 30–31, and 1972*b*, pp 435–36)

[11] There are, of course, polar cases in which relative prices do not change and in which changes in relative prices do not affect the distribution of income Qualifying remarks of this type are not repeated in the discussion that follows

We can use some simple algebra to learn more about this outcome. First, the expression for the supply curve in quadrant a is $dQ_x/Q_x = S(dP_x - dP_y)$, as noted above. The shifted demand curve can be written as follows.

$$dQ_x/Q_x = -E(dP_x - dP_y) + dQ_G/Q_x, \qquad (1a)$$

where dQ_G/Q_x is the fractional amount by which demand shifts toward good x because of public procurement, at the initially prevailing prices Solving these two equations simultaneously yields the following.

$$dP_x - dP_y = dQ_G/Q_x(S + E) \qquad (9)$$

and

$$dQ_x/Q_x = S\,dQ_G/Q_x(S + E) \qquad (9a)$$

The more elastic is either supply or demand, the less the price of x rises relative to the price of y as demand shifts toward good x Similarly, the more price elastic is demand, the less the quantity of good x rises as demand shifts to good x But the more elastic is supply the more output responds to the shift in demand

Figure 1 can be used to derive the same result First, note that OM in figure 1 equals $OW\ S$ and MF equals $OW \cdot E$ But $OM + MF = OF$ Thus, $OF = OW\ (S + E)$, and $OW = OF/(S + E)$ But because OW is simply $dP_x - dP_y$ and OF is dQ_G/Q_x or $(m_g - m_p)\ dG/Q_x$, this yields the results reported in equation (9) Similarly, $OM = S\ OF/(S + E)$, as reported in equation (9a) in slightly different terminology

Quadrant a can tell us the incidence on the side of uses of income of a shift in purchases toward product x But to learn about incidence on the side of sources of income it is necessary to "go behind" the supply curve to learn how relative factor prices are affected This can easily be done, of course, since the supply curve was derived from the more fundamental relationships described in quadrants b–d In particular, for the policy in question we need only note that the indicated increase in output of good x implies an inflow of capital of ON, which in turn raises the wage rate in x by the fraction OV_x By the same token, the wage rate in industry y falls

---

[12] A second curve could be added to quadrant b to show how any change in capital utilization is reflected in the output of good y through the following equation·

$$dQ_y/Q_y = -g_K K_x\ dK_x/K_x K_y \qquad (2a)$$

This equation does not, of course, figure in the final solution, except to give the change in output of good y.

by OV_y, as capital flows from that industry.[13] Finally, though strictly speaking the question is of no interest from the standpoint of incidence analysis, since in that context only changes in relative product prices and relative factor prices are relevant, we might want to solve for the changes in the prices of goods x and y, given that $dP_K = 0$.[14] This we can do in quadrant d; the price of good x rises by OW_x and that of good y falls by OW_y. Of course, W_yW_x equals OW in quadrant a by construction.

IV. Tax Incidence

The model presented here accommodates analysis of taxes levied on nine distinct tax bases: value added, retail sales, or income (1), only one of the two products (2), the services of only one of the two factors in all its uses (2), and the services of one of the two factors in one of the two industries (4). But in the neoclassical world of this model we need examine explicitly the incidence of only four of the nine taxes. The incidence results for the rest of the nine follow directly by subtraction.[15] Yet it will be useful for the present purpose to examine the incidence of more than the minimum number of taxes. In particular, we will examine the incidence of the general sales or income tax, the tax on product x, the tax on both uses of the mobile factor, the tax levied on the mobile factor in one industry, and the tax levied on the immobile factor in one industry. In each case we assume that tax proceeds are used to finance a distributionally neutral increase in public spending. This last assumption is relaxed in Section V, in which we consider the possibility of using these taxes to finance an increase in public spending that is not neutral.

---

[13] The results given in equations (9) and (9a) and equations (2)–(4) can be used to solve for these capital flows and price changes. Thus

$$dK_x/K_x = S\,dQ_G/Q_xf_K(S + E), \tag{9b}$$

$$dP_{Lx} = S\,dQ_G/Q_x\sigma_xf_K(S + E), \tag{9c}$$

and

$$dP_{Ly} = SK_x\,dQ_G/Q_xK_y\sigma_yf_K(S + E). \tag{9d}$$

These results, like those of equations (9) and (9a), though different in appearance, are the same as those reported in the mathematical treatment in McLure (1972*b*, p. 444). The same approach can be employed to solve for the changes in absolute product prices, given $dP_K = 0$.

[14] One reason for doing this is that we might wish to choose an alternative numeraire, say good y. To convert percentage changes in a given price under the numeraire assumption $dP_K = 0$ to percentage changes under an assumption of $dP_y = 0$ is a simple matter. One need only subtract from each percentage price change as originally calculated (including dP_K) the calculated value of dP_y. An alternative and attractive approach is to formulate the diagrammatic analysis presented here under the assumption of $dP_y = 0$. Space constraints dictate that this alternative formulation be left as an exercise for the interested reader.

[15] See Musgrave (1959, pp. 348–54) and McLure (1971*b*, p. 29).

The supply curve developed in Section II and used in Section III to examine expenditure incidence is based upon a no-tax situation To be able to allow for the influence of various kinds of taxes on the supply curve, we must modify equations (3)-(6) Equations (3) and (4) become

$$\frac{dK_x}{K_x} - \frac{dL_x}{L_x} = -\sigma_x(dP^*_{Kx} - dP^*_{Lx}) \tag{3a}$$

and

$$\frac{dK_y}{K_y} - \frac{dL_y}{L_y} = -\sigma_y(dP^*_{Ky} - dP_{Ly}) \tag{4a}$$

In these equations (and in other equations below), asterisked terms refer to changes in various factor prices inclusive of factor taxes That is, they refer to the changes in factor prices that affect entrepreneural decisions and costs of products They are to be distinguished from their unasterisked relatives, which indicate changes in factor prices net of factor taxes, that is, what factor owners receive The following relations between factor prices gross and net of factor taxes will be used in this section [16]

$$dP^*_{Kx} = dP_K + dT_{Kx} + dT_K, \tag{10a}$$

$$dP^*_{Lx} = dP_{Lx} + dT_{Lx}, \tag{10b}$$

and

$$dP^*_{Ky} = dP_K + dT_K \tag{10c}$$

Note that while capitalists receive the same net price for capital services in the two industries, entrepreneurs will pay different gross prices for capital services if capital is taxed at different rates in the two industries Thus the asterisked value of the change in the price of capital carries a sectoral subscript

Equations (5) and (6) must be rewritten as follows to allow for the existence of excise taxes, general consumption taxes, and taxes on factors

$$dP_x = f_L\, dP^*_{Lx} + f_K\, dP^*_{Kx} + dT_x + dT_c \tag{5a}$$

and

$$dP_y = g_L\, dP_{Ly} + g_K\, dP^*_{Ky} + dT_y + dT_c \tag{6a}$$

These equations now state that the percentage change in the price of a product is the weighted (by initial shares) average of the percentage change in the price of the two factors, including any factor taxes, plus any ad valorem tax levied on the product itself or upon consumption

We can begin the analysis with the tax on the output of industry x Such a tax affects only two curves in the analysis, the curve P_x in quadrant d relating the price of good x to the wage rate in x and (through that) curve

---

[16] Equations (3a), (4a), and (10a)-(10c) reflect the choice of taxes to be analyzed, dP_{Ly} appears without an asterisk because no tax on labor in y is to be examined

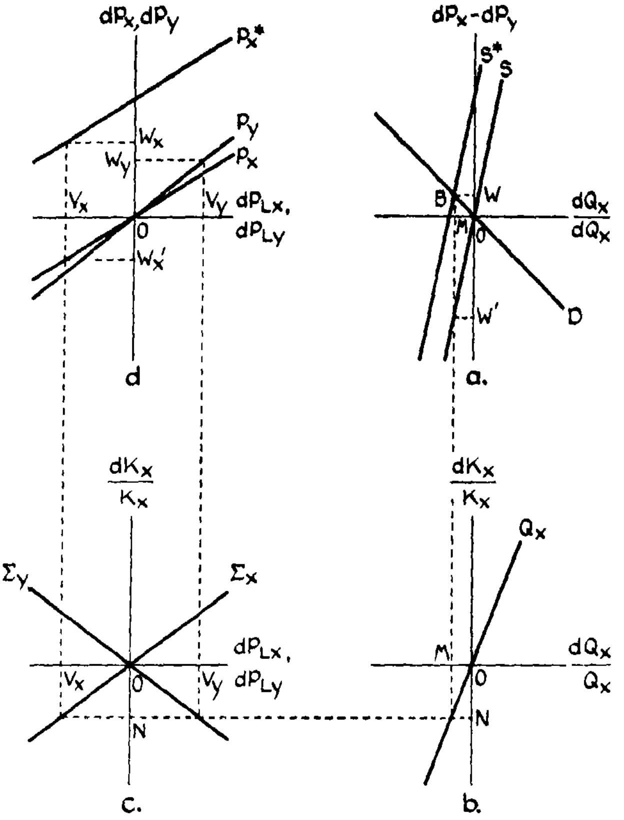


Fig 2 —Analysis of incidence of excise tax on good *x*

S, the supply curve of good *x* It moves both these curves vertically upward by the amount of the tax to P_x^* and S^*, as shown in figure 2[17] The intersection of the shifted supply curve S^* and the demand curve *D* shows that the tax increases the relative price of good *x* by *OW* and that the output of the taxed good falls by *OM* Carrying the output effect to quadrant *b*, we see that the tax induces a fractional capital outflow from the taxed industry of *ON*. The reallocation of capital results in a fall in the wage rate (given $dP_K = 0$) in the taxed industry of OV_x and a rise in the wage rate in the untaxed industry of OV_y, as shown in quadrant *c*. Finally, we can use these effects and the curves in quadrant *d*, as before, to determine the effects on the prices of the two goods The price of good *x* rises by

---

[17] It is a simple exercise in the manipulation of the diagrams to show that the upward shift in curve P_x in quadrant *d* to P_x^* shifts the supply curve by the same amount.

OW_x and that of good y rises by OW_y, the difference W_yW_x being the change in relative product prices, which is equal to OW in quadrant *a*. Total payments to factors in industry x per unit of output fall in nominal terms by OW_x'. By construction, $W_x'W_y$ in quadrant *d* equals OW' in quadrant *a*, the amount by which the price of good x, net of tax, falls relative to the price of good y.

As in the case of our examination of expenditure incidence, we can use some simple algebra to elaborate on these results. First, the supply curve in this case can be written as

$$dQ_x/Q_x = S(dP_x - dT_x - dP_y), \tag{7a}$$

since firms react to prices net of tax. Combining this equation with equation (1) above, we have

$$dP_x - dP_y = \frac{S}{E + S} dT_x \tag{11}$$

and

$$dQ_x/Q_x = \frac{-ES}{E + S} dT_x \tag{11a}$$

By the same reasoning used in the previous section, we can easily derive these results from figure 2. $OB = OM + MB$, which in turn equals $-(E + S)OW$. But OB also equals $-S\,dT_x$. Thus, $OW = S\,dT_x/(E + S)$, as reported in equation (11). Finally, $OM = -ES\,dT_x/(E + S)$, as reported in equation (11a). As partial equilibrium analysis would suggest, the increase in the relative price of the taxed good induced by excise taxation depends positively upon the elasticity of supply and negatively upon the elasticity of demand.[18] Changes in factor rewards can be calculated directly from these results by the use of equations (2)–(4), and those in absolute product prices by also using equations (5a) and (6a). Or, in the spirit of the present analysis, they can be read from the appropriate curves in quadrants *c* and *d* of figure 2.

The next tax to be examined, a general consumption or income tax, is particularly easy to analyze. Such a tax can be thought of as an ad valorem tax levied at the same rate on both goods x and y. Thus it raises both curve P_x and curve P_y in quadrant *d* by the amount of the tax, as can be seen by inspection of equations (5a) and (6a). But because both these curves rise by the same amount, the tax has no effect on the supply curve for industry x. Thus it has no effect on relative product prices, the

---

[18] Musgrave (1959, pp. 291–92) notes the analogous partial equilibrium result. It is worth noting at this point that the expressions in equations (1) and (1a) are each 1 degree higher in E and S than are those in equations (9) and (9a). Because of the nonuniform wedge driven between costs and prices by excise taxes that is not present when government procurement simply shifts the demand curve, more substitution in production and consumption is involved in adjusting to excises than to expenditure shifts. This is noted as well in McLure (1972*a*).

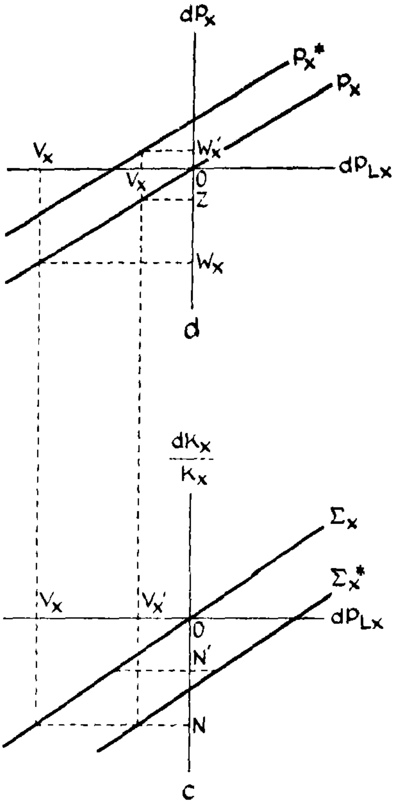


FIG 3 —Analysis of incidence of tax on capital in industry x

interindustry allocation of capital, or the relative rewards to factors It does, however, given the numeraire equation $dP_K = 0$, raise the absolute prices of both goods x and y by the amount of the tax And, what is more crucial, it raises product prices relative to factor prices by the ad valorem amount of the tax Thus it is simply borne in proportion to initial shares in national income (or consumption)

The analysis of a tax levied on capital services in only industry x is more complicated than those presented thus far First, the tax shifts curve Σ_x in quadrant c to the right by the amount of the tax (to Σ_x^* in fig 3), as can be seen from equations (3a) and (10a) In addition, it raises curve P_x in quadrant d by f_K times the amount of the tax, as shown by equations (5a) and (10a), to P_x^* To determine the effect of the tax on the supply curve, we can choose an arbitrary point N on the vertical axis of quadrant c in figure 3 In the absence of the tax, the capital outflow indicated by ON would result in a fall in the wage rate in industry x of

OV_x, a fall in the price of good x of OW_x, and some (unspecified) rise in the wage rate and product price in industry y. But given the tax on capital in x and the choice of capital as numeraire, this capital outflow would result in a fall in the wage rate in industry x of only OV_x', as shown by curve Σ_x^*. Similarly, the impact of the postulated capital flow on the price of x, given the tax, can be seen from curve P_x^* in quadrant d to be OW_x'.

Since the existence of the tax on capital services in industry x does not affect the way the wage rate and product price in industry y respond to the postulated capital inflow, the supply curve is shifted upward by the tax on capital used in x by the amount W_xW_x'. But this shift can readily be seen to be exactly equal to the ad valorem rate of tax on capital services used in industry x, as follows. First, W_xW_x' can be decomposed into two parts, W_xZ and ZW_x'. The first of these is simply equal to $f_L\, dT_{Kx}$ (The distance V_xV_x' equals dT_{Kx}, as noted above, and the slope of curve P_x is f_L) The other is the distance between curves P_x and P_x^*, noted above to be $f_K\, dT_{Kx}$. Since $f_K + f_L = 1$, the entire distance W_xW_x' equals the ad valorem rate of taxation of capital in industry x, dT_{Kx} This implies, of course, that this tax raises the supply curve of x by exactly as much as an equal ad valorem excise tax on good x This, in turn, implies that with one exception the tax on capital services in industry x affects all variables in the model exactly as the excise tax does The exception is the wage rate in industry x.

The effect the tax has upon the wage rate in x is given in quadrant c Suppose that the tax on capital services in x produces the outflow of capital from the industry of ON shown in figure 3 In such a case the wage rate falls by OV_x' But this result of a lower wage rate in industry x cannot be generalized Suppose that production in industry x is characterized by a high elasticity of factor substitution, relative to the elasticity of product demand. In that case it is possible for the wage rate in that industry to rise relative to the net price of capital services This case is shown in figure 3 for the capital outflow ON' The point is that the capital outflow reduces the output of good x and raises its price as well as reducing the marginal physical product of labor If demand is sufficiently price inelastic and factor substitution in industry x sufficiently elastic, the former influence can outweigh the latter. In that case the tax on capital in x will actually raise the wage rate in that industry, relative to the return on capital[19]

The final two taxes to be considered are a tax on labor in industry x and a tax on capital in both industries These taxes are considered together because both are borne entirely by the owners of the taxed factors. The tax on labor in industry x shifts two curves. Σ_x in quadrant c and P_x in quadrant d. But because both these curves shift to the left by exactly the amount of the tax, the supply curve in quadrant *a* is unaffected, and the tax has no effect on relative product prices or the allocation of capital between industries Nor does it affect the relative returns to capital and labor in the untaxed industry Its only effect is to reduce the net wage rate in the taxed industry by the amount of the tax. Thus labor specific to one industry bears the full burden of any tax levied on it. Since the wage rate gross of tax in industry x and the wage rate in industry y are both unchanged, the absolute price of neither good changes

---

[19] If the numeraire equation $dP_y = 0$ is chosen, the sign of dP_{Lx} depends simply upon ($\sigma_x - E$) (see McLure 1971*b*, p. 39).

From this result it is easy to see why the tax on capital services in x and the excise tax on x have the same effects, except on the wage rate, if levied at the same ad valorem rate The excise is equivalent to an ad valorem tax on both factors used in x But the part of the excise that is equivalent to an ad valorem tax on labor in x affects only the net wage rate in that industry Thus in total the excise and the tax on capital affect all other product and factor prices identically In terms of the diagrams, the part of the excise that is equivalent to a tax on capital in x shifts Σ_x to the right, but the part that is equivalent to a tax on labor in x shifts it to the left by an equal amount, leaving the curve unchanged on balance The two taxes together raise curve P_x by dT, the amount of the ad valorem tax But so does the tax on capital services in industry x, which, however, also affects Σ_x

The previous result was the standard Ricardian result that a tax on an immobile factor is borne by that factor A similar, but less widely recognized, result occurs if a tax is imposed on all uses of the mobile factor capital.[20] Such a tax shifts curves Σ_x and Σ_y in quadrant *c* to the right by the amount of the tax, and it raises curves P_x and P_y by $f_K\, dT_K$ and $g_K\, dT_K$, respectively But it has no effect on the supply curve for good x This can be demonstrated using figure 4 and the reasoning employed in the discussion of figure 3 Alternatively, we can simply use the results for a tax levied on capital services in one industry, since a general tax on capital services is equivalent to a tax levied on capital services at the same rate in each industry The part of the general capital tax that is equivalent to a tax on capital in x shifts the supply curve upward by the ad valorem amount of the tax, the part equivalent to a tax on capital in y offsets this shift Yet another way to see that the supply curve is unaffected is to recall that a tax on capital in one industry has the same impact on the supply curve as an equal ad valorem excise on the output of that industry But we have already shown that an excise tax levied at the same ad valorem rate on both industries does not affect the supply curve In any event, relative product prices and the interindustry allocation of capital are unaffected by the general tax on capital services But the wage rates in

---

[20] There is no need to examine explicitly the incidence of a tax on labor in both industries. Such a tax can be broken down conceptually into two taxes—one on labor in industry x and one on labor in industry y. Since each of the parts would be borne by the taxed labor, the entire tax is borne by labor

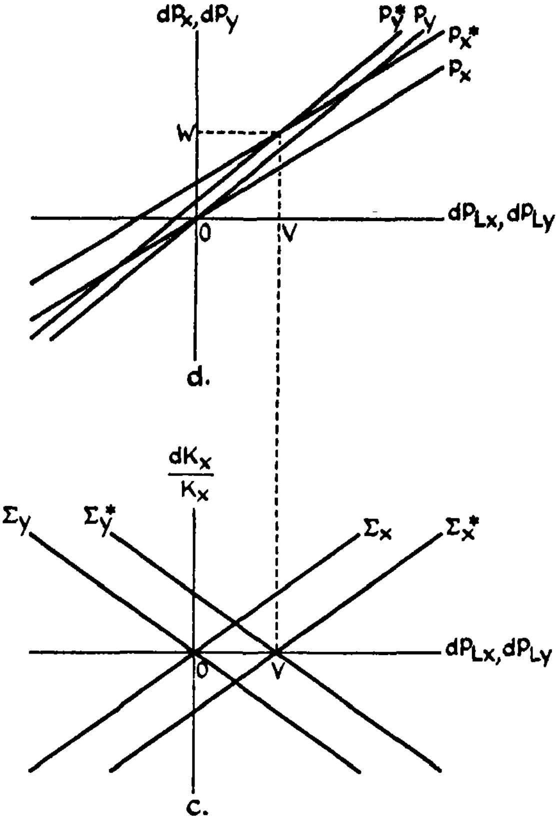


FIG 4 —Analysis of incidence of tax on all capital income

the two industries and the prices of the two goods rise by the ad valorem amount of the tax, given the choice of capital as the numeraire In terms of figure 4, the solution is *OV* in quadrant *c* and *OW* in quadrant *d* That *OW* equals dT_K can be shown by the same reasoning employed above in the discussion of figure 3 In real terms this implies that owners of capital bear the full burden of the universal tax on capital even though capital is mobile between industries.

V. Balanced-Budget Incidence

Thus far we have considered the incidence of government expenditures financed in a distributionally neutral way (Section III) and the incidence of several taxes used to finance distributionally neutral expenditures (Section IV). Now we consider briefly how the analysis of the last two sections can be combined to allow us to investigate the incidence of balanced-budget increases in government expenditures in the general

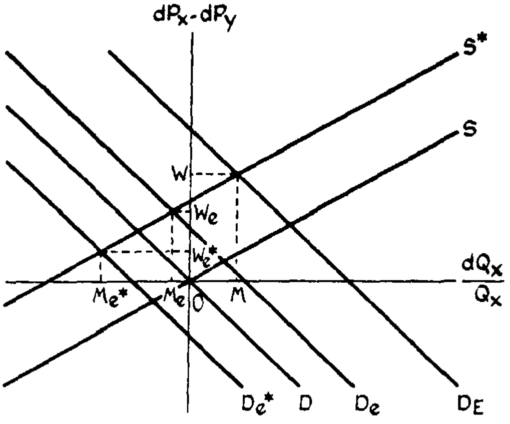


FIG 5 —Analysis of balanced-budget incidence

case in which neither expenditures nor their financing is distributionally neutral

We need to examine only the effect of financing a nonneutral increase in public expenditures through a tax on good x There is no need to examine the cases of financing through taxes on either the immobile factor or the mobile factor in all its uses Since those taxes are simply borne by the taxed factors, they can readily be combined with nonneutral expenditure patterns And since the tax on good x and the tax on capital in x have different results only for the wage rate in x, there is no need to consider the case of balanced-budget financing through a tax on capital in x. Finally, we can limit ourselves to quadrant a, since all that is of real analytical interest occurs there, though to obtain the full solution for all the variables in the model, it would be necessary to carry the results from quadrant a into the remaining quadrants, as we have done in earlier sections [21]

The diagrammatic analysis is given in figure 5, which combines aspects of quadrant a in figures 1 and 2 The demand curve shifts from D to D_E because of the public expenditure policy, and the supply curve shifts from S to S^* because of the tax on good x [22] The new equilibrium is determined by the intersection of the two shifted curves, D_g and S^*. In this particular case the price of good x would rise relative to that of good y by OW and output of good x would rise by the fraction OM Of course, if the shift in aggregate demand toward good x were less than shown by curve D_g, the price rise would be less than shown, and if the demand shift were small enough, output of good x would fall, rather than rise Such a case is shown by curve D_e. Finally, if demand shifts away from good x, rather than toward it, as it does in the case of curve D_e^*, the relative price of good x would rise less than if the tax financed a neutral expenditure program, and it might even fall

---

[21] Curve P_x in quadrant d is the only curve in quadrants b–d that is affected by the balanced-budget program financed by a tax on good x It rises by the amount of the tax, as demonstrated in Section IV Similarly, for each of the taxes not explicitly considered in this section, the shifts of curves in quadrants b–d are exactly those described in Section IV. In each case the demand curve shifts as described in Section III

[22] Careful examination of fig 5 and eqq. (12) and (12a) will reveal that strictly speaking curves D_E and S^* are not consistent, in that D_E should intersect the horizontal axis no further to the right of the origin than S^* intersects the vertical axis above the origin, if $m_y \leq 1$ and $m_x \geq 0$. The exaggeration in fig 5 allows the possible outcomes described in the text to be shown more clearly than would otherwise easily be the case.

We can use equations (1a) and (7a) to solve explicitly for the changes in relative prices induced by balanced-budget expenditures [23] Alternatively (and equivalently), we can simply combine the results given in equations (9) and (9a) and equations (11) and (11a) In either event, we reformulate the fractional shift in demand for good x resulting from the increase in government expenditures as follows

$$dQ_G/Q_x = (m_g - m_p)\, dG/Q_x, \tag{12}$$

where dG is the balanced-budget increase in spending, and m_g and m_p are the public and private marginal propensities to spend on good x Moreover, we note that because of the assumption of a balanced-budget fiscal operations, the following holds

$$dG = Q_x\, dT_x \tag{12a}$$

Using these relations gives us the following solution for the balanced-budget increase in spending under consideration

$$dP_x - dP_y = (S + m_g - m_p)\, dT_x/(S + E) \tag{13}$$

and

$$dQ_x/Q_x = -S[E - (m_g - m_p)]\, dT_x/(S + E) \tag{13a}$$

Of course, if $m_g = m_p$, these results reduce to equations (11) and (11a), as they must

The approach developed in earlier sections can be employed to demonstrate these results diagrammatically, though dealing with a balanced-budget change is slightly more complicated than treating either the tax or expenditure sides by itself In figure 6, $MB = -S\ OW$ and $MF = E \cdot OW$ Thus $BF = OW \cdot (E + S)$. But BF also equals $(S + m_g - m_p)\, dT_x$. From this we can see that $OW = (S + m_g - m_p)\, dT_x/(E + S)$, as stated in equation (13). Finally, to solve for OM, the fractional change in the

---

[23] The results reported in this section are consistent with those in McLure (1972b, pp. 447–50).

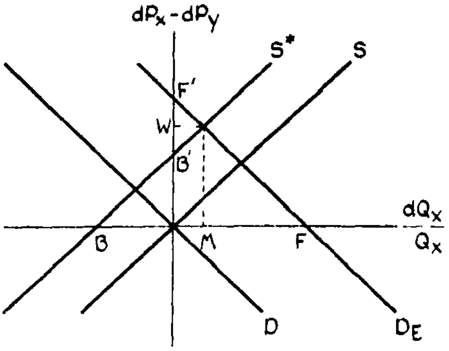


Fig. 6.—Analysis of balanced-budget incidence.

output of good x, we need only note that $B'W = OM/S$ and $WF' = OM/E$, so that $B'F' = OM(E + S)/ES$ and $OM = B'F' \cdot ES/(E + S)$. But we can rewrite $B'F'$ as $(m_g - m_p) \, dG/Q_x \cdot E - dT_x$. Thus $OM = S(m_g - m_p - E) \, dT_x/(E + S)$, as noted in equation (13a).

Inspection of figure 6 reveals that the price of good x can rise by more than the amount of tax on good x only if demand shifts toward that good enough so that output of that good rises on balance. Equation (13a) shows that this result can occur only if $(m_g - m_p)$ exceeds E. Of course, equation (13) gives this result directly. Finally, we can see that the price of good x can fall relative to that of good y even if a tax on good x is employed to finance a balanced-budget increase in public spending if supply is relatively price inelastic and demand shifts strongly toward good y. In particular, $dP_x - dP_y$ has the sign of $S + m_g - m_p$, which is negative for relatively small values of S and m_g and large values of m_p.

VI. Discrete versus Small Changes

Strictly speaking, the Harberger model, as well as its many applications and extensions, is applicable only to the analysis of small taxes and expenditures, as it is written in terms of differentials. Yet it is commonly utilized as the basis for policy decisions involving tax changes that are far from small. The appropriateness of the model for such uses has been the subject of increasing skepticism (see Krauss 1972, for example). In this section we examine briefly one of the problems posed by using the Harberger analysis of differential changes to approximate the effects of discrete tax and expenditure changes. But the problem is only exposited. No attempt is made either to generalize the model to deal adequately with discrete changes or to perform a sensitivity analysis of the inaccuracy resulting from using the Harberger model to deal with discrete changes.

Those more demanding tasks are beyond the scope of this brief discussion (but see Shoven and Whalley 1972).

In a nutshell, the problem is that to be fully consistent with the Harberger model as usually written, all the curves in figures 1–6 should be straight lines.[24] But it is virtually certain that for discrete changes in taxes or expenditures no more than two of the curves will in fact be straight lines. If such is the case, the Harberger model—and its various extensions and applications—can be expected to yield quantitatively incorrect results.

Whether curves D and Σ_x should be straight lines does not concern us. If the income-compensated elasticity of product demand with respect to relative product prices is not constant, then D should not have a constant slope. And if the elasticity of factor substitution is not constant, then Σ_x will not be a straight line But these are empirical questions which, while important, are not central to the discussion of this section. Rather, the crucial questions are whether curves Q_x, Σ_y, P_x, P_y, and S are straight lines. These questions can be answered on analytical grounds

The slope of curve Q_x in quadrant b is the inverse of the elasticity of output of good x with respect to inputs of capital. In general, this elasticity is not constant for increasing doses of capital applied to a constant amount of labor. Only if the elasticity of factor substitution is unity is the output elasticity constant If (and only if) the elasticity of factor substitution is less than unity, as seems likely, the output elasticity falls as more and more capital flows into industry x Thus curve Q_x should increase in slope as capital flows into industry x (Moreover, the increase in the slope is greater, the smaller is the elasticity of factor substitution) In qualitative terms, the curve relating capital inflows to changes in output should look like Q_x^*, rather than Q_x (see fig 7)

Moving to quadrant c, we see at once that curve Σ_y cannot generally have a constant slope. Since that slope is $-\sigma_y K_y/K_x$, it falls in absolute value as capital flows into industry x, even it we assume that σ_y is a constant Of course, if K_y/K_x is initially quite large, curve Σ_y is virtually

---

[24] Each of equations (1)–(6) describes a straight line One problem not considered in this paragraph is that in the discrete case the demand curve is likely to shift. For differentially small tax policies it is reasonable to assume for analytical convenience that marginal propensities to consume good x are equal for all economic agents But for discrete tax changes this assumption becomes untenable even as an analytical device, since in the limit marginal propensities are equal to average propensities. Thus, let us simply note that redistribution of discrete size is likely to shift the demand curve, but in an undetermined way. Mieszkowski (1967, pp 261–62) has analyzed the impact of relaxing the assumption that all expenditure patterns are identical at the margin. While this problem, which is closely related to the question of expenditure incidence, is easily treated algebraically, it would be difficult to incorporate it in the present diagrammatic analysis. For a potentially more damaging criticism, in which it is claimed that excess burdens of taxation makes the assumption of constant real income untenable, see Ballentine and Eris (1973).

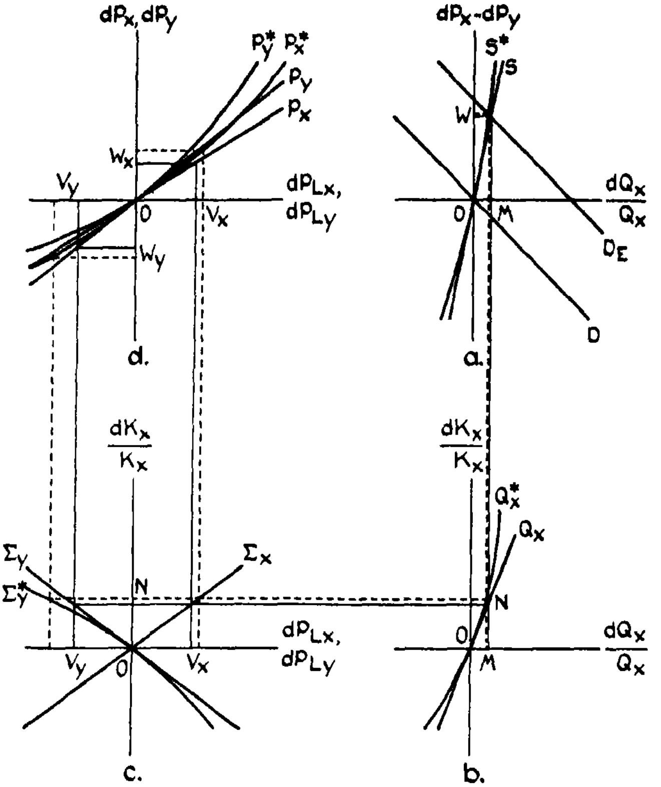


FIG 7.—Illustrative analysis of expenditure incidence under linear and nonlinear versions of the Harberger model

vertical, and its slope is not much affected by even proportionally large inflows of capital into industry x (i.e., by large values of dK_x/K_x). But if K_y/K_x is initially small, Σ_y will be relatively flat, and even small positive values of dK_x/K_x will flatten it further by an appreciable amount.[25] In qualitative terms, we can expect the appropriate curve to resemble Σ_y^* in figure 7, instead of the linear form Σ_y.

Finally, curves P_x and P_y in quadrant d have slopes equal to the initial shares of labor in the two industries. By the same reasoning that prevailed in quadrant b, we know that these slopes are constant only if the elasticity of factor substitution is unity. But if factor substitution is less than unitary elastic, labor's share in either industry will increase with its price, and in qualitative terms the curves will look like P_x^* and P_y^*, rather than P_x and P_y, in figure 7.

---

[25] For example, if K_y/K_x were initially 1/2, a capital flow of $dK_x/K_x = 1/2$ would result in a value of K_y/K_x of 0 and a horizontal Σ_y curve.

Turning now to quadrant *a* we can see that the supply curve is likely to bend upward to the right. This can be seen from an inspection of either the other three quadrants of figure 7 or equation (8). If both elasticities of factor substitution are less than one, an increase in output of good x is accompanied by a fall in f_K and a rise in f_L, a rise in K_x/K_y, and a fall in g_L.[26] Only the fall in g_L tends to swing the supply curve S (derived from the unasterisked or straight-line curves in quadrants *b–d*) downward All the other changes tend to swing it upward From this it seems reasonable to conclude that in general supply becomes less elastic as we move from the origin to the right in quadrant *a* of figure 7, and more elastic as we move to the left, though the results clearly depend upon the values of the parameters of the model [27]

The importance of this qualification can be seen by considering the expenditure policy indicated by the shift of the demand curve from D to D_E in figure 7. The linear version of the model gives the results indicated by the solid line connecting points corresponding to OM and OW, ON, OV_x, and OV_y, OW_x and OW_y, reading around the diagram in a clockwise direction from quadrant *a*. The curvilinear model, on the other hand, gives the results indicated in the same notation, but corresponding to the points connected by a broken line [28] Comparing the two results, we find that the linear model understates the change in relative product and factor prices and overstates the quantity adjustment induced by the expenditure policy [29] Quantitative estimates of the extent to which linearization in the Harberger model causes results to be inaccurately stated is beyond the scope of this paper

In a similar vein, we could introduce an excise tax into the system to see how the linearization might lead us astray But again, doing so would lead us further into the quanitative evaluation of errors resulting from linearization than we should go after a paper has reached this length. Suffice it to say that under certain circumstances the errors can be substantial and that users of the linear version of the Harberger model should keep this caveat in mind in offering policy advice and in making empirical estimates on the basis of the linear model.

---

[26] If both these elasticities are greater than one, most of what is said here is reversed If only the elasticity of substitution in industry y exceeds unity, supply is certain to be less elastic as output increases than in the linear version If the elasticity exceeds unity only in industry x, supply is likely to be more elastic as output increases

[27] Differentiating equation (8) totally helps one to see the roles played by the various parameters Unfortunately, the total derivative is not unambiguous in sign It may be worth noting explicitly that whereas P_x/P_y can increase without bound, it is bounded on the bottom by $P_x/P_y = 0$

[28] Since the primary interest here is in qualitative results, no effort has been made to draw curve S^* to correspond exactly to curves Q_x^*, Σ_y^*, Σ_x, P_x^*, and P_y^*. Unfortunately, even fairly casual treatment of this subject tends to plunge us directly into the sensitivity analysis that is beyond the scope of this paper

[29] We have a familiar index number problem here The results reported here would be reversed if the slopes derived from the posttax, rather than the pretax, equilibrium were employed in the comparison. A similar point is noted in Mieszkowski (1969, p. 1109).

References

Ballentine, J. Gregory, and Eris, Ibrahim "On the General Equilibrium Analysis of Tax Incidence." Discussion Paper no. 38, Rice Univ., Program of Development Studies, 1973.

Harberger, Arnold C. "The Incidence of the Corporation Income Tax." *J.P.E.* 70 (June 1962)· 215–40.

——— "Efficiency Effects of Taxes on Income from Capital." In *Effects of Corporation Income Tax*, edited by Marian Krzyzaniak. Detroit: Wayne State Univ. Press, 1966

Krauss, Melvyn "Differential Tax Incidence: Large versus Small Tax Changes." *J P E* 80 (January/February 1972)· 193–97.

McLure, Charles E , Jr "Interregional Incidence of General Regional Taxes." *Public Finance* 24 (1969) 457–83

——— "Taxation, Substitution, and Industrial Location." *J P E.* 78 (January/February 1970). 112–32 (*a*)

——— "Tax Incidence, Macroeconomic Policy, and Absolute Prices " *Q J.E.* 84 (May 1970) 254–67 (*b*)

———. "The Design of Regional Tax Incentives for Colombia." In *Fiscal Reform for Colombia*, edited by Malcolm Gillis. Cambridge, Mass.· Harvard Law School, 1971 (*a*)

———. "The Theory of Tax Incidence with Imperfect Factor Mobility." *Finanzarchiv* 30 (1971) 27–48. (*b*)

———. "General Equilibrium Incidence Analysis· The Harberger Model after Ten Years " Discussion Paper no. 37, Rice Univ , Program of Development Studies, 1972. (*a*)

——— "The Theory of Expenditure Incidence " *Finanzarchiv* 30 (1972) 432–53 (*b*)

Mieszkowski, Peter M "On the Theory of Tax Incidence " *J.P.E* 75 (June 1967) 250–62

———. "Tax Incidence Theory: The Effects of Taxes on the Distribution of Income." *J Econ Literature* 7 (December 1969) 1103–24

Musgrave, Richard A *The Theory of Public Finance.* New York McGraw-Hill, 1959

Shoven, John B , and Whalley, John "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U S " *J. Public Econ* 1 (November 1972) 281–321.

An Economic Theory of Suicide

Daniel S. Hamermesh

Michigan State University

Neal M. Soss

Bureau of the Budget, State of New York

> . as soon as the terrors of life reach the point at which they outweigh the terrors of death, a man will put an end to his life [A SCHOPENHAUER, *On Suicide*]

Although sociologists have developed and tested numerous theories about suicide, economists have not analyzed this phenomenon We derive an economic theory of suicide and test its implications using (1) data by age in many developed countries, (2) a time series, 1947–67, by age group in the United States, (3) a cross section by state and age group in 1960 Most of our predictions are verified Particularly striking are the response to unemployment, as strong in the postwar period as in the Depression and stronger among older individuals, and the negative effect of increased permanent incomes among all but the youngest age group, a result found in both the time series and the cross section

Since the appearance of Durkheim's *Le suicide* in 1897, sociologists have constructed numerous theories to explain patterns in suicide rates both within and across societies Economists have not considered the problem of suicide, although it is surely one involving individual decision making

The work for this paper was done while both authors were at Princeton University. They thank Joshua Greene for his highly competent research assistance. Ronald Ehrenberg, Ray Fair, Robert Goldfarb, Simon Rottenberg, Burton Weisbrod, and the referees provided helpful comments on an earlier draft Participants in seminars at Princeton University, Michigan State University, and the National Bureau of Economic Research also offered useful suggestions.

and often also economic factors.[1] Just as economics has served to improve our understanding of fertility (cf. Becker 1960; Schultz 1969) and marriage (Silver 1965), so it may provide a testable theory of suicide. The empirical analysis should enable us to measure another effect on society of both long-run economic growth and cyclical fluctuations in income. In this paper we outline several sociological theories and then develop a new, economic theory of suicide. Some implications of this theory are then tested against various sets of data.

I. Sociological Theories and an Economic Theory

Most subsequent work by sociologists in the study of suicide has been motivated by Durkheim In addition to providing a model for empirical research by sociologists, he summarized existing knowledge about suicide patterns by age, literacy, and other demographic and sociological variables and constructed a typology for classifying suicides In a departure from the Durkheimian model, Henry and Short (1954) developed the frustration-aggression theory According to them, suicide results from individuals' frustration in their attempts to achieve their social goals and the ensuing aggressive feelings, directed either at themselves (suicide) or at others (homicide) They postulate that economic improvement leads to a decrease in frustration and thus aggression, and they concentrate most of their empirical work on the relation of suicide rates to the business cycle They also argue that frustration and aggression occur differentially depending upon the degree of external restraint imposed upon the individual Since they assume that external restraint is inversely related to status, their theory implies a positive correlation between suicide rates and status Their empirical work demonstrates conclusively only the positive correlation of suicides and unemployment in the United States before 1942.

Undoubtedly, much of the variation in suicide rates among ethnic and demographic groups can be explained by these and other sociological theories Several aspects of this problem, particularly variations in suicides by age and income, that cannot simultaneously be reconciled by any one of these theories may well be rationalized by an economic theory. Admittedly, some suicidal behavior can in no way be attributed to economic factors We contend, though, that some is due to economic decision making, so that some of the variation in suicide rates should be explicable using hypotheses derived from an economic theory[2]

---

[1] In the earliest available records in Great Britain, covering 1770 through 1830, one-fourth of the recorded suicides were attributed to poverty (Fedden 1938, p. 203). Lendrum (1933) analyzes 419 male suicide attempts and finds that over one-third of those giving a motive cite economic problems. Shneidman and Farberow (1965, p. 36) analyze a sample of suicide notes and find 8 percent of those listing a specific cause mention poverty or business difficulties.

Let us postulate the following utility function for the average individual in a group of people with permanent income YP·

$$U_m = U[C(m, YP) - K(m)] > 0, \tag{1}$$

where m is his age and K a technological relation describing the cost each period of maintaining himself alive at some minimum level of subsistence[3] (Presumably $K' > 0$) If this is the utility of the average individual age m with permanent income YP, then the present value of his expected lifetime utility at age a is

$$Z(a, YP) = \int_a^{\omega} e^{-r(m-a)} U_m P(m)\, dm, \tag{2}$$

where r is the private discount rate, ω is the highest attainable age, and $P(m)$ is the probability of survival to age m given survival to age a The Z, then, is a decreasing function of a and an increasing function of YP

Let $b_i \sim N(0, \sigma^2)$ be the ith individual's taste for living, or conversely, his distaste for suicide The variable b_i is in units commensurate with the utility function and is subject to any transformations performed on U_m[4] It is assumed to be defined for the cohort at birth

Following the spirit of the epigraph to this paper, we assume that an individual kills himself when the total discounted lifetime utility remaining to him reaches zero This point must be a maximum, for the remaining lifetime utility is decreasing with age Our hypothesis, therefore, is that the ith individual commits suicide if and when $Z_i(a, YP) + b_i = 0$ The fraction of individuals in the cohort born at time $(t - a)$ who commit suicide at age a is

$$S(a) = f[-Z(a, YP)], \tag{3}$$

where $f(\cdot)$ is the density function for b_i[5] In our model, the instantaneous suicide rate is simply the fraction of individuals in the cohort for whom

---

[2] This is not the first analysis of the economics of mortality However, previous work neither formulates theories nor tests hypotheses Hansen (1957) presents a calculation of the loss in GNP resulting from child mortality, Holtmann and Ridker (1965) compute the welfare loss arising from the early incurrence of burial costs of individuals who die prematurely We calculate that the loss in human capital resulting from suicides among males ages 25–64 in 1959 was $740 million, based on a discount rate of 6 percent (*Census of Population* [1960], PC(2)-7B, and U S Office of Vital Statistics, *Vital Statistics* [1959])

[3] Grossman (1972) has argued that this function can be affected by the individual's earlier choice of health expenditures

[4] Our results are valid for any distribution which is positive over the entire real number line and has positive first and second derivatives at the left tail

[5] We are explicitly assuming that no members of the birth cohort die from other causes before age a. Relaxing this assumption by dividing equation (3) by the fraction of the cohort actually remaining at age a changes none of our theoretical results on the relation of suicide to age and income

$Z(a, YP)$ reaches $-b$ at age a. Differentiating equation (3) totally:

$$dS = -f'\left(\frac{\partial Z}{\partial a}\, da + \frac{\partial Z}{\partial YP}\, dYP\right). \tag{4}$$

Now $\partial Z/\partial a < 0$, both because $K' > 0$ and because there is less gross consumption left to enjoy as a person ages. Also, $\partial Z/\partial YP > 0$, so we can use equation (4) to derive the predictions that the suicide rate, S, will rise with age and decrease with permanent income

Taking the total differential of equation (4), we obtain.

$$\begin{aligned} d^2S = &\left[-f'\frac{\partial^2 Z}{\partial a^2} + f''\left(\frac{\partial Z}{\partial a}\right)^2\right] da^2 \\ &+ 2\left[-f'\frac{\partial^2 Z}{\partial YP\, \partial a} + f''\frac{\partial Z}{\partial YP}\cdot\frac{\partial Z}{\partial a}\right] dYP\, da \\ &+ \left[-f'\frac{\partial^2 Z}{\partial YP^2} + f''\left(\frac{\partial Z}{\partial YP}\right)^2\right] dYP^2 \end{aligned} \tag{5}$$

Of the three terms, only the sign of the third can be determined unambiguously It is positive, given the reasonable assumptions that we are dealing with the left tail of the normal distribution (so that f' and $f'' > 0$) and that the marginal utility of lifetime income decreases with increases in YP so that $\partial^2 Z/\partial YP^2 < 0$ The signs of the other terms are ambiguous The first term in $\partial^2 S/\partial a^2$ is negative if Z is decreasing in a, and the second is positive Similarly, the first term in $\partial^2 S/(\partial YP\, \partial a)$ is positive and the second negative [6]

The age-specific suicide rates, then, will depend upon the distribution of tastes against suicide and the distribution of permanent incomes Several specific hypotheses arise from the economic theory we have outlined We expect that the rate of suicide increases with age ($\partial S/\partial a > 0$); that the suicide rate is inversely related to permanent income ($\partial S/\partial YP < 0$), but that the marginal absolute effect on suicide declines as permanent income increases ($\partial^2 S/\partial YP^2 > 0$) [7]

---

[6] The analysis could be expanded to include expected utility based on the probability of employment at each age Unless severe restrictions are placed on differences by ages in the elasticities of expectations about future unemployment in response to changes in current unemployment, the only result is that suicide rates will rise as unemployment increases. In Section IV we examine the role of unemployment and, by inference, differences by age in these elasticities

[7] In addition to the economic aspects of suicide discussed here, some demographic differences in suicide rates are well established The rate is lower for females than males in all countries compiling reliable data, and in the United States the nonwhite suicide rate is slightly less than half that of whites (see U.S. Office of Vital Statistics, *Vital Statistics* [1967], II.a: 1.20) These differences are subsumed under the taste variable, b, in our theory, an assumption that allows us to concentrate on economic factors. A sociological theory that performs well in explaining these demographic differences is that of Gibbs and Martin (1964).

II. Empirical Evidence—Introductory Considerations

The deficiencies in the data on suicide rates are well known. (Douglas [1967] presents a detailed discussion of this problem) Underreporting of suicide rates varies in different jurisdictions depending upon the persistence of the coroner and the degree to which society stigmatizes suicide as a form of deviant behavior. This difficulty implies that any cross-country comparisons of suicide rates are likely to be worthless, the underreporting problem is undoubtedly one reason for the low suicide rates observed in predominantly Catholic countries [8] Within the United States, the use of suicide rates by states in a cross-section regression model could be dangerous because errors in the dependent variable are correlated with the true (observed) values of the income variables Errors may also be introduced into any correlation of income and suicide by the requirement that an individual be covered for 2 years before an insurance company will pay off on a death by suicide (Gregg 1959, pp 139–40).

While we admit that there are problems in using official statistics on suicide rates, a consideration of them suggests that the data should be useful if handled judiciously Data on suicide rates by age within any country should embody very few problems, there is little reason to expect systematic differences in underreporting of suicides by age The degree of underreporting may change over time as the moral stigma attached to suicide decreases We can, however, account for this difficulty in Section IV and still make use of time-series data on suicide rates Data from a narrow jurisdiction in which one official has charge of all decisions on causes of death should be fairly reliable Finally, the use of the appropriate demographic measures should enable us in Section V to remove most of the errors arising from comparisons across jurisdictions within a fairly homogeneous society

III. Suicide by Age—International Evidence

Table 1 presents the average male suicide rate by age in each of the 21 developed countries for which all the data are available from 1965 through 1967 In nine of these, the rate rises monotonically with age, in two it reaches a peak at the oldest age group although it does not rise steadily, and in eight others it rises monotonically to a peak in the age group 55–64 Only in two countries, Sweden and Poland, does the peak occur before this point, and in the former the rate is essentially constant after the peak While the evidence presented in this table is by no means conclusive, it does suggest that within a country there is a general tendency

---

[8] This problem is vividly illustrated by the probably apocryphal quotation, cited by Alvarez (1972, p. 87), from an Irish coroner who certified an accidental death from gunshot wounds. "Sure, he was only cleaning the muzzle of the gun with his tongue."

TABLE 1

MALE SUICIDE RATES BY AGE (PER 100,000 POPULATION), 21 COUNTRIES, 1965–67 AVERAGE

COUNTRY	AGE					
	15–24	25–34	35–44	45–54	55–64	65–74
Australia	10.1	22.4	29.7	33.9	34.9	41.8
Austria	19.7	31.6	42.3	50.0	65.0	59.5
Belgium	7.8	13.0	17.8	29.5	42.1	54.9
Bulgaria	8.3	7.9	10.9	19.1	25.2	40.2
Canada	10.3	15.9	19.4	25.5	30.0	25.9
Czechoslovakia	26.0	35.9	43.6	50.2	52.7	67.3
Denmark	9.7	22.5	32.6	45.9	49.7	40.2
England and Wales	6.4	10.8	14.5	17.7	23.9	26.8
Finland	15.5	35.2	53.2	62.2	68.6	68.9
France	7.0	17.0	26.0	40.4	54.3	54.0
W. Germany	18.3	26.9	33.0	46.1	55.9	48.4
Hungary	30.5	44.1	56.0	65.2	70.7	76.4
Israel	4.8	10.2	11.4	16.4	18.2	23.0
Japan	14.7	21.0	15.5	20.9	35.2	52.1
Netherlands	4.1	7.5	8.4	13.9	18.8	24.5
New Zealand	6.7	11.0	19.4	29.0	29.9	28.3
Norway	6.2	10.5	15.6	19.5	20.6	19.8
Poland	13.0	22.6	29.2	30.4	28.5	24.4
Sweden	13.8	28.3	40.8	51.0	50.5	47.4
Switzerland	19.0	25.1	34.5	46.2	49.3	53.0
United States	9.9	17.3	22.4	28.4	36.0	35.7

SOURCE.—World Health Organization, *World Health Statistics Annual*, 1965, 1966, 1967 (Geneva: WHO, 1968–70).

for suicide rates to rise with age. Insofar as people born earlier have both lower incomes at any age and less income remaining, this evidence is consistent with our theory of the effect of income on suicide.

IV. Time Series—Males, United States, 1947–67

Data are available on the suicide rate by 5-year age groups for 1947–67, and the data on real incomes of persons exist for some closed 5-year and 10-year age intervals through age 64 in the United States.[9] These data form the basis for the empirical analysis in this section. The theoretical model we presented in Section I provides us with all but two of the variables to be used in the regression equations. In particular, for a cohort age A at time t, the suicide rate $S(A, t)$ will be a function of the discounted real permanent income stream remaining at time t and the age of the cohort. We also include a variable to measure transitory deviations around expected income. When unemployment rises, individuals' expectations of future incomes (and utilities) are revised downward. Holding real income of the employed constant, an increased number of people will believe future prospects to have diminished and will commit suicide

---

[9] The suicide data, numbers of suicides by 5-year age groups, are published in U.S. Office of Vital Statistics, *Vital Statistics* (1967) and earlier issues. These were transformed into rates per 100,000 population through the use of population data kindly made available to us by the National Center for Health Statistics. Data on money incomes of male persons by age for 1947–64 are available in Bureau of the Census, *Trends in the Income of Families and Persons in the United States, 1947–64* (Washington: Government Printing Office, 1967). For 1965–67 the data were provided by Charles Beach of Queen's University and are based on his calculations using published data. The money income series are deflated by the Consumer Price Index (CPI). The CPI series and data on unemployment rates by age group are from Bureau of Labor Statistics, *Handbook of Labor Statistics* (1971), Bulletin 1705, pp. 120 and 253. All the income data used are deflated to constant 1967 dollars.

We construct discounted permanent income remaining, $YPL(A, t)$ by assuming that individuals project the average growth rate of incomes by age group that occurred during 1947–67 We divide each 10-year age interval in two and assign each 5-year interval the value $Y(A, t)$, the mean real income at age A at time t, for the 10-year interval to which it belongs Through a regression of this variable against time we construct $\bar{Y}(A, t)$, expected real income at age A at time t, which we project out to the end of the economic lifetime of each group

We then compute $YPL(A, t)$ as:

$$YPL(A, t) = \bar{Y}(A, t) + \sum_{i=1}^{L} \frac{\bar{Y}(A + 5i, t + 5i)}{(1 + r)^{5i}} + \frac{\bar{Y}(N, t + N - A)}{(1 + r)^{N-A}},$$

where N is the mean age of males in the open interval 65+, and $L = (62.5 - A)/5$[10] Age A is taken as the midpoint of the appropriate 5-year age interval The series is computed for $r =$ 02, 06, 12, 20, and ∞.

For transitory income we use the proxy $UN(A, t)$, the fraction of males age A who are unemployed at time t This term is included alone and in an interaction with the age variable There is evidence that periods of unemployment are longer among older workers who are laid off, and further evidence that older workers' expectations of finding new employment are affected by this phenomenon.[11] We thus expect unemployment to have an increasing effect on suicide as workers age[12]

---

[10] The actual value of N in 1970 was 73.6 (Computed from Bureau of the Census, *Census of Population* [1970], General Population Characteristics, PC(1)-B1, table 50) This method of computing permanent income remaining implicitly assumes that individuals are uncertain about their incomes in the distant future and weight them accordingly. The implication of this uncertainty is an optimal consumption path which is shaped like a monotonic function of $\bar{Y}$ as we have computed it (see Nagatani 1972 for a derivation of this optimization rule).

[11] Wilcock and Franke (1963, p 61) find that only 40 percent of workers below age 45 failed to find jobs within one year after their layoffs, for workers 45 and over the figure was 60 percent. One older worker claimed, "It's little better than shooting me to throw me out of a job at 55" (p. 85).

[12] An additional transitory term, $Y(A, t) - \bar{Y}(A, t)$, was included in preliminary work. It was dropped because the t-statistic of its coefficient never exceeded .2 in absolute value.

Our basic estimating equation is:

$$S(A, t) = \alpha_0 + \alpha_1 YPL(A, t) + \alpha_2 YPL^2(A, t) + \alpha_3 A + \alpha_4 A^2 + \alpha_5 A^3 + \alpha_6 UN(A, t) + \alpha_7 UN(A, t) \cdot A + v_{A,t}, \tag{6}$$

where $v_{A,t}$ is a disturbance term, and income is measured in thousand-dollar units. The higher-order terms on A are included as a result of some experiments made on the international data in table 1.[13] We estimate equation (6) over a pooled cross section of the time series 1947–67 for each of the nine 5-year age intervals, 20–64. The open interval 65+ is excluded because most males in the group are not economically active. A modified version of this equation, denoted (6′), is also estimated and differs from equation (6) only by the addition of the term $\alpha_8 t$. This modification enables us to test the hypothesis that the apparent effect of increasing income over time on suicide rates merely reflects a trend in the latter. Finally, we search over the YPL series computed for each of the five discount rates and report results only for the value that maximizes the likelihood function in this parameter.

Examination of the residuals based on ordinary least-squares estimates of equations (6) and (6′) showed that $E(v_{A,t} \cdot v_{A,t-1}) \neq 0$ for each A. To remove this source of inefficiency in the estimation, we devised a two-round procedure that transformed the original variables by an estimate of the serial correlation of the residuals within each age group.[14] The results of this procedure are presented in table 2 for $r = 20$, the value of the discount rate that maximized the likelihood function for both equations (6) and (6′). This fairly high discount rate suggests that incomes in the near future are most important in affecting suicide. The R^2 presented is based on the second-round estimates.[15] Undoubtedly the most striking

---

[13] In 10 of the 20 countries in table 1 for which suicide rates generally increase with age, the suicide-age profile has a definite logistic shape. To reflect this phenomenon in equation (6), and thus to distinguish better between K and YP in testing our theory, we need the higher-order terms.

[14] No standard technique for removing serial correlation in pooled cross-section and time-series data is available. To circumvent this difficulty we employed the following method. We postulate a variance-covariance matrix of the residuals composed of nine 21 × 21 blocks on the prime diagonal

$$\Omega = \sigma^2 \begin{pmatrix} 1\,\rho\,\rho^2 & & \rho^{20} \\ \rho & & \\ \cdot & & \\ \rho^{20} & \cdot\ \cdot & 1 \end{pmatrix},$$

and blocks consisting entirely of zeros off the prime diagonal. The first-round estimate $\hat{\rho}$ is obtained, and each variable is transformed to produce $\Delta X_{A,t} = X_{A,t} - \hat{\rho} X_{A,t-1}$, $t = 1948, \ldots, 1967$. The second-round estimates are based on the ΔX variables for the sample 1948–67.

[15] There is no problem of nondiagonality of the variance-covariance matrix of the residuals resulting from correlation among the residuals within each age group. Only 2 percent of the total unexplained variance in equations (6) and (6′) is accounted for by this correlation.

TABLE 2
SECOND-ROUND ESTIMATES OF EQUATIONS (6) AND (6′), $r = .20$

Variable	(6)	(6′)
Constant	26.54	26.24
	(5.29)	(3.89)
YPL	−7.73	−7.78
	(−5.82)	(−4.10)
YPL^2	56	57
	(6.20)	(5.54)
A	−2.08	−2.09
	(−2.86)	(−2.69)
A^2	.065	065
	(3.73)	(3.56)
A^3	−.0005	−.0005
	(−3.62)	(−3.47)
UN	−70.16	−69.68
	(−3.15)	(−3.12)
UN A	3.40	3.38
	(5.68)	(5.64)
t		.008
		(.06)
R^2	93	93
ρ	517	526

NOTE.—*t*-values in parentheses here and in table 3

TABLE 3
MEAN VALUES OF DERIVATIVES AT DIFFERENT AGES, BASED ON EQUATION (6)

DERIVATIVES	GROUP		
	20–24	40–44	60–64
$\partial S/\partial UN$	6.23	74.14	142.04
	(.57)	(7.79)	(7.63)
$\partial S/\partial Y$	2.17	−.29	−1.11
	(10.62)	(−1.62)	(−4.02)
$\partial^2 S/\partial Y^2$	3.16	3.12	1.45
	(6.20)	(6.20)	(6.20)

feature of these results is the insignificance of the trend variable, *t*, in equation (6′), apparently the effects of income on the suicide rate depicted in equation (6) are not the result of the correlation between income and time. In both equations all the other variables are significant, and those for which we postulated an expected sign on the coefficient confirm our expectation The estimates for the age terms suggest that in the United States there is a logistic shape curve relating suicide rates to age.

In table 3 we present various derivatives based on equation (6) and computed at the means for each age group listed. (We assume the means have no stochastic element and treat them as constant in computing variances of these derivatives.) Raising the fraction unemployed by .03, roughly equal to the average increase for males in postwar recessions, increases the suicide rate by amounts ranging from 0.19 per 100,000 for the youngest age interval to 4.26 per 100,000 for the 60–64 age group. In absolute numbers, we find that total male suicides per year in the age groups 20–64 rise by approximately 950 in an average recession, given the population distribution of the United States in 1960. Durkheim (1951) and others have demonstrated the relation between suicide and unemployment in periods characterized by violent swings in business activity. Our results indicate that this relation also holds for the relatively mild recessions of the postwar period and that it is stronger among higher age groups.[16]

Table 3 also presents the derivatives involving income that we discussed in Section I. The effect of a thousand-dollar increase in all real expected incomes on the suicide rate is negative for the median and oldest age groups, but positive for the 20–24-year-old group. We find that $\partial S/\partial Y = 0$ at $A = 35$, so that we observe the predicted sign of this derivative in all but the three youngest age groups.[17] The predictions of our model on $\partial^2 S/\partial Y^2$ are confirmed by the signs of this derivative.

As incomes have risen in the postwar period, the male suicide rate has fallen for most age groups, especially for the older groups nearing retirement age. Only for the youngest three age groups (20–34 years) has the suicide rate risen with rising incomes, a result that is contrary to our predictions. There is no simple economic explanation for this phenomenon, but the expansion in the number of people pursuing education into their twenties and the consequent postponement of consumption by an increased fraction of the people in these groups could be one cause. The decline in suicide rates for older people as their incomes have increased may be especially strong because of the decrease in variability of income resulting from the expansion of Social Security benefits. Indeed, there was a sharp drop in the relative suicide rates of groups over age 55 in the late 1930s coincident with the introduction of Social Security.

V. Cross-Section Evidence

In this section we present first a summary of studies on the relationship of suicide rates to occupation and income based on individual data and then the results of estimating the cross-section analogue to equation (6). Official published reports contain no information relating the suicide

---

[16] Our results imply an average peak-to-trough change in the male suicide rate for ages 20–64 of 2.07. Interestingly enough, the weighted suicide rate for this group rose in the Depression by 10.08, from a low of 28.59 in 1928 to a peak of 38.67 in 1932, while the fraction unemployed rose by 192. Our estimates of equation (6) imply a rise of 13.27 in this suicide rate in response to this change in the fraction unemployed, a result fairly close to that which occurred.

[17] Equations (6) and (6′) were reestimated for each age group separately, without the three age terms and for $r = 20$. In all nine groups the net effect of UN was positive; furthermore, $\partial S/\partial Y$, computed at the mean for each age group and based on the coefficients for that group alone, was again positive for ages 20–34 and negative thereafter.

TABLE 4
RELATIVE SUICIDE RATES (100 = PROFESSIONAL RATE) AND FAMILY INCOMES IN 1959, BY OCCUPATION*

	COOK COUNTY (1959–63)†		NEW ORLEANS (1954–59)‡		TULSA (1937–56)§	
OCCUPATION	Males	Total Y	White Males	Whites' Y	White Males	Whites' Y
Professionals	100	9,445	100	8,456	100	8,522
Managers	107	10,487	174	8,669		
Clericals	88	7,086	100	6,020	33	6,782
Sales	134	8,843	102	6,861		
Craftsmen	141	8,142	137	6,262	41	6,201
Operatives	157	6,887	224	5,527	58	5,359
Service	314	6,045	502	4,689	68	4,108
Laborers	342	6,026	481	4,276	110	4,316

* Computed from *Census of Population* (1960), PC(1)-15D, 20D, 38D, table 145, for the SMSA containing the city

† Computed from Maris (1967, p 249)

‡ Computed from Breed (1963, p 181)

§ Computed from Powell (1958, p 135)

rate to income or occupation Several sociologists have, however, constructed detailed data by occupation for individuals within a narrow jurisdiction, and the results of these studies are summarized in table 4 along with census data on family incomes by occupation of the family head In each study the suicide rate for professional workers is used as a base of 100, and the index for each occupation group is related to it In all three studies presented, the suicide rate is generally lower among the occupation groups in which income is greater As table 5 shows, the rank correlations of the suicide index with income are negative, although the small sample size makes it difficult to claim a high degree of significance for them individually The correlations are especially strong when professionals and managers are deleted from the sample The data presented in table 4 do not hold constant for religious practices, if there is a negative correlation between the percentage Catholic and occupational status within each narrow jurisdiction, the results presented in table 4 will be biased against observing the expected negative relation between suicide and occupational status While there are relatively few Catholics in Tulsa, the fairly high percentage in New Orleans and Cook County suggests that the true pattern of suicide rates declines even more sharply with income Age differences by occupation would impart a negative bias to the correlations The correlations of the suicide index and median age by occupation in each city are all found to be nearly zero, suggesting that differences in age by occupation are not producing spurious correlations in table 5.

Dublin (1963), using data from England and Wales, presents results similar to those that Powell (1958) has produced for Tulsa. Sainsbury

TABLE 5

RANK CORRELATIONS BETWEEN SUICIDE AND INCOME BASED ON TABLE 4

	Cook County	New Orleans	Tulsa
All occupations	− 69*	− 63*	− 31
Excluding professionals and managers	− 83*	− 77*	− 90*

* Significantly different from zero at the 10% level, one sided test

(1955) finds constant suicide rates except in the highest and lowest social strata, where the rates are greater, but he, like Dublin (1963), fails to adjust for age differences Summarizing these studies, we can conclude that, with the exception of professional workers and managers, a strong negative relation between income and suicide rates seems to be present in the data The more general conclusion, that the suicide rate declines monotonically as permanent income increases, is supported somewhat less strongly by the available studies

We turn now to the last test of our theory of suicide Whereas the units of observation in Section IV are various age groups at different points in time, the units of observation here are age groups in different states of the United States at the same point in time, 1960 We restrict our analysis to the white male population and to all 38 states in the contiguous United States for which data on income by age and race are available[18] The age and income variables are constructed similarly to those used in the time-series tests in Section IV Five age groups—20–24 and the four 10-year groups between 25 and 64—form the basis for analysis To avoid heteroskedasticity in the residuals, each observation is weighted by the square root of the white male population for that age and state

Special problems in the cross-section test require the introduction of two new variables[19] The first difficulty is that, *ceteris paribus*, suicide rates tend to be substantially higher in the Far West The restless life often associated with this area may be a cause of this phenomenon This consideration suggests the use of a dummy variable, assuming the value of unity for California, Oregon, and Washington The second is that the distribution of Roman Catholics in the population is not random across states Rather, Catholics tend to be concentrated in the heavily urban, industrial states where incomes are relatively high Given the lower incidence of suicide among Catholics, a failure to control for their proportion in a state's population would bias the coefficient of income negatively We might then spuriously confirm the implications of our theory for the effects of income on the suicide rate. To correct for this possibility, we experimented with several possible indices of the proportion of a state's population who are Catholics, including the fraction of foreign stock and the fraction with Spanish surname The variable best justified theoretically and used here is $CATH$, the percentage of total elementary and secondary enrollment in Catholic parochial schools in each state This proxy may be better than a measure of the percentage Catholic population, if, as in Ryder and Westoff's study of fertility (1971), Catholic education is the best indicator of the degree to which behavior conforms to Church dogma[20]

---

[18] The states excluded are Idaho, Maine, Montana, Nevada, New Hampshire, North Dakota, Rhode Island, Utah, Vermont, and Wyoming

[19] These problems can be interpreted in the context of our theory as resulting from differences in the distributions of *b* across different cultures or societies

Table 6 presents the estimates of

$$\begin{aligned} S(A,j) = \beta_0 + \beta_1 YPL(A,j) + \beta_2 YPL^2(A,j) + \beta_3 A + \beta_4 A^2 \\ + \beta_5 A^3 + \beta_6 CATH + \beta_7 WEST + \mu_{A,j}, \end{aligned} \tag{7}$$

where $S(A,j)$ is the suicide rate by age and state, and the YPL and A terms are as used in equations (6) and (6′)[21] Equation (7′) differs only in that the observations for the far western states and the dummy variable $WEST$ have been deleted Table 7 presents estimates of the derivatives $\partial S/\partial \bar{Y}$, based on equation (7), computed as in table 3.

The first thing to note about these tables is that, as in the time-series tests, the predictions of our theory on the age profile of suicide rates and the effect of income on suicide rates are borne out by the data The sign patterns of $\partial S/\partial \bar{Y}$ by age in the cross-section and time-series tests are remarkably alike—negative for all but the youngest age groups The only difference is for the 25–34 group, where the derivative is positive in the time series but negative here The positive value of $\partial^2 S/\partial \bar{Y}^2$ here is contrary to our prediction in Section I Given that the search procedure maximizes at $r = \infty$ and given the age-YPL profile resulting from this, the signs on YPL and YPL^2 are the only ones consistent with the pattern

---

[20] Numbers of white male suicides by state for 1960 are published in U.S Office of Vital Statistics, *Vital Statistics* (1960), table 9-6 These were transformed into rates per 100,000 population using *Census of Population* (1960), data published in PC(1)-1B, table 59 Income data by age group for each state are published in PC(1)-D, table 134 Data on Catholic elementary enrollment in 1961–62 are from Office of Education, *Circular* no 753, p 34, those on secondary enrollment, available only for 1960–61, are from *Circular* no 707, p 27, and public school enrollment for 1961–62 is from *Circular* no 757, p 47

[21] Initial estimates, using the same five values of r as in Section IV, showed that the likelihood function reaches a maximum at $r = \infty$ We then searched over a finer grid of $r = 35$, 50, 75, and 1 00 and found $r = \infty$ still to be the maximizing value, although the likelihood function is very flat for $r \geq 20$ No unemployment term is included in the estimates of equation (7) that we present because of our lack of confidence in the state unemployment data Moreover, it is not clear that differences in unemployment rates across states at a point in time represent transitory components of income When the measured unemployed rate, prorated by the national unemployment rate for the age group relative to the total national rate, was added to equation (7), its coefficient had a t-value of $-$ 6 in equation (7), its addition produces only minor changes in the other coefficients

TABLE 6

ESTIMATES OF EQUATIONS (7) AND (7′), $r = \infty$

Variable	(7)—38 States	(7′)—35 States
Constant	46.45 (2.01)	43.98 (1.88)
YPL	10.97 (3.35)	12.84 (3.69)
YPL^2	−1.33 (−3.86)	−1.61 (−4.31)
A	−4.80 (−2.42)	−4.88 (−2.43)
A^2	.14 (3.02)	.14 (3.06)
A^3	−.001 (−3.16)	−.001 (3.26)
CATH	−.43 (−5.75)	−.39 (−5.24)
WEST	7.96 (5.18)	.
R^2	.893	.867

NOTE.—*t*-values in parentheses here and in table 7.

TABLE 7

MEAN VALUES OF $\partial S/\partial\bar{Y}$ IN THE FIVE AGE GROUPS

Group	20–24	25–34	35–44	45–54	55–64
(7)	3.89 (2.50)	−2.02 (−2.47)	−3.61 (−3.64)	−2.51 (−2.95)	−.33 (−.40)
(7′)	4.31 (2.65)	−2.68 (−3.15)	−4.53 (−4.33)	−3.21 (−3.61)	−.57 (−.66)

of $\partial S/\partial\bar{Y}$ by age observed both here and in the time series. That the likelihood function is maximized at $r = \infty$ in the cross section is consistent with the observation that income differences across states are permanent. Therefore, current incomes are a good proxy for lifetime incomes here and may well perform better.[22]

The *WEST* dummy is positive and significant; its coefficient indicates that, once we adjust for income, age, and Catholic school enrollment, the suicide rate for males 20–64 averages eight more per 100,000 in the Far West. The proxy variable for the extent of Catholicism, *CATH*, has the expected negative coefficient. The percentage Catholic enrollment varies from 1 in North Carolina to 26 in Wisconsin; it implies a difference in the adult male suicide rate of 10 to 11 per 100,000 population, if other factors are held constant. Since the average rate in our sample is only 25 per 100,000, the effect of Catholicism is quite substantial.

---

[22] We are indebted to Jacob Mincer for this observation. The flatness of the likelihood function in the cross section for $r \geq 20$ suggests that there is little difference across states among these measures.

VI. Conclusion

In no sense do we claim that the individual agony resulting in suicide stems from what is solely an economic calculation, the majority of suicides can perhaps be explained on noneconomic grounds. Nonetheless, as Fedden (1938, p. 54) pointed out, poverty can be an important cause of suicide in developed societies where values are to a large extent based on the possession of material goods. Given the materialistic basis of modern society, it is reasonable to expect that variations in the suicide rate will be related to economic variables in ways predictable by economic theory.

Although our main purpose has been to demonstrate the value of applying economic theory to the problem of suicide, we have also discovered some interesting and apparently novel empirical results. Within the United States we have found that the mild cyclical decreases in economic activity since 1945 have produced the same proportionate increases in suicide rates that resulted from the stronger cyclical fluctuations before 1945 The suicide behavior of older people is significantly more sensitive to variations in unemployment than is that of younger people Suicide rates decrease with increased incomes, both in the cross section and the time series, for all but the youngest age groups Contrary to widely held beliefs (Shneidman and Farberow 1957, pp 60–67) based on studies using inappropriate data or inadequate analysis and on what we call the "Richard Cory myth," our results imply that suicide rates are generally lower among higher-income groups

Our results suggest the value of using economic theory to analyze social phenomena such as suicide Like other theories, ours has successfully predicted the relation of suicide by age, however, ours also entails predictions about both the response of suicide to increased income and the change in that response as income increases that seem to be confirmed, with some exceptions, by the data It is more convincing when one remembers the trend toward decreased underreporting of suicides that should bias our time-series results against confirmation of our theory We believe the study demonstrates the need for work using data on individuals to separate out economic influences on patterns of suicide from noneconomic ones

References

Alvarez, A. *The Savage God* New York. Random House, 1972.

Becker, G. "An Economic Analysis of Fertility." In National Bureau of Economic Research, *Demographic and Economic Change in Developed Countries*. Princeton, N.J.: Princeton Univ. Press, 1960.

Breed, W. "Occupational Mobility and Suicide among White Males." *American Soc. Rev.* 28, no. 2 (April 1963): 179–88.

Douglas, J. *The Social Meaning of Suicide.* Princeton, N.J.: Princeton Univ. Press, 1967.

Dublin, L. *Suicide: A Sociological and Statistical Study.* New York: Ronald, 1963.

Durkheim, E. *Suicide.* Translated by J. Spaulding and G. Simpson. 1897. Reprinted Glencoe, Ill.. Free Press, 1951.

Fedden, H. R. *Suicide· A Social and Historical Study.* London· Peter Davies, 1938.

Gibbs, J. P, and Martin, W. T *Status Integration and Suicide.* Eugene: Univ. Oregon Press, 1964.

Gregg, D. W *Life and Health Insurance Handbook.* Homewood, Ill.: Irwin, 1959.

Grossman, M "On the Concept of Health Capital and the Demand for Health." *J.P.E* 80, no 2 (March-April 1972) 223-55

Hansen, W. L. "A Note on the Cost of Children's Mortality." *J.P.E.* 65, no. 3 (June 1957) 257-62.

Henry, A F., and Short, J F *Suicide and Homicide.* Glencoe, Ill.. Free Press, 1954.

Holtmann, A, and Ridker, R. "Burial Costs and Premature Death" *J.P.E* 73, no. 3 (June 1965) 284-86

Lendrum, F C "A Thousand Cases of Attempted Suicide." *American J. Psychiatry* 13, no 3 (November 1933) 479-500

Maris, R "Suicide, Status and Mobility in Chicago." *Soc. Forces* 46, no 2 (December 1967) 247-56

Nagatani, K. "Life Cycle Saving Theory and Fact." *A E R.* 62, no. 3 (June 1972)· 344-53.

Powell, E. "Occupation, Status and Suicide" *American Soc. Rev.* 23, no. 2 (April 1958) 131-39

Ryder, N, and Westoff, C *Reproduction in the United States, 1965.* Princeton, N J.: Princeton Univ Press, 1971

Sainsbury, P. *Suicide in London* London· Chapman & Hall, 1955

Schultz, T. P "An Economic Model of Family Planning and Fertility." *J.P.E* 77, no 2 (March-April 1969) 153-80

Shneidman, E S, and Farberow, N *Clues to Suicide* New York McGraw-Hill, 1957

——— *The Cry for Help* New York McGraw-Hill, 1965

Silver, M "Births, Marriages, and Business Cycles in the United States" *J P E* 73, no. 3 (June 1965) 237-55

Wilcock, R, and Franke, W *Unwanted Workers* Glencoe, Ill. Free Press, 1963.

John Maynard Keynes: Scientist or Politician?

Elizabeth Johnson

Royal Economic Society

Keynes thought of himself as a social scientist, an expert adviser to government on economic problems, yet he possessed—and used—all the talents of a natural born politician. His upbringing and position in society somewhat circumscribed his thinking, but when he was unable to solve the problem of mass unemployment with the orthodox tools at hand, his scientific integrity and his flexibility of mind impelled him to make the imaginative leap forward of the General Theory Despite the limitations of his time and place, he was truly a scientific political economist.

John Maynard Keynes—scientist or politician? The reader of the popular press of a generation ago would have had no doubt of the answer. Keynes, a swinging weather vane of a man, was the most unscientific of individuals —a cartoonist's dream He was Keynes the india-rubber man· the *Daily News and Chronicle* of March 16, 1931, carried an article headed "Economic Acrobatics of Mr. Keynes" and illustrated it by a sketch of "A Remarkable Performance Mr John Maynard Keynes as the 'boneless man' turns his back on himself and swallows a draught"—the draught, a glass marked "15 per cent Protection." After years of preaching the virtues of free trade, he had first announced the end of laissez faire and now urged a revenue tariff on the country. As an exasperated political opponent remarked on another occasion—complaining of the man who in 1925 said wage costs were too high and in 1929 wanted higher prices: "It is difficult to reconcile

---

This paper was read at the annual meetings of the British Association for the Advancement of Science, Section F, at Leicester, September 5, 1972.

Mr. Keynes the politician with Professor Keynes the economist. He seems to be both right and wrong!"[1]

Keynes himself had no such difficulty. In his own opinion, he was always right. He had a clear idea of his own role in the world; he was the economist—at first Cassandra, croaking prophecies of doom about the economic consequences of reparations and the gold standard, prophecies which came all too true—and then as he gathered stature, the chief economic adviser to the world, to the chancellor of the exchequer of the day, to the French minister of finance whoever he was, to the president of the United States To elaborate, Keynes the economist initially thought of himself as the educator, the persuader, the man who would assemble all the relevant information and thereby start the reverberation of public opinion that would echo back to the politicians who, he said, "have ears but no eyes" (Keynes 1922*b* Hereafter all references to Keynes's writings are by earliest publication date) Then as he became established as an expert, he came to think of himself more as the economic scientist, the technician, the mechanic who is called in to fix the machine when the self-starter is broken He looked forward to the time when economists would be consulted like dentists and hailed President Roosevelt as the first head of state to take theoretical advice as the basis for large-scale action.

"For the next 25 years, in my belief," he wrote in 1932, "economists, at present the most incompetent, will be nevertheless the most important, group of scientists in the world And it is to be hoped—if they are successful—that after that they will never be important again" (1932*a*)

Keynes had a generally low opinion of politicians as charlatans who manipulated the public with their propaganda and obstinately clung to the accepted shibboleths until the winds of change forced them to tack. He knew himself to be an intellectual and a scientist, but he was a very *political* economist, addressing himself to the big problems of his time As the high tide of the nineteenth century ebbed away and the waves of the twentieth came rolling in, as the world struggled out from the aftermath of the 1914–18 war with its old antagonisms and old sovereignties, the new hopes of peace and progress were bogged down by old debts and old habits. Poverty stood in the midst of what should have been plenty, and in Britain, less hard hit than Germany or the United States, one quarter of the working population was unemployed How was the world to get out of this mess? In Britain, what could be left to the individual and what must be done by the state? No longer would "private ambition and compound interest . . . between them carry us to paradise" (1927), the system could no longer be trusted to correct itself. The problems were political problems and for Keynes they were intellectual ones

---

[1] The remark was made by Sir Laming Worthington-Evans, Conservative secretary of state for war, in a letter to the *Evening Standard*, May 6, 1929, part of a published exchange with Keynes at the time of the 1929 election campaign. During this election Keynes not only supported but took the lead in originating the Lloyd George program to remedy unemployment.

For all that, he was a natural politician He inherited the Liberal politics of his parents and at Cambridge was a Liberal president of the union. During the early part of his life, he often appeared on the platform to speak for Liberal candidates He was asked to stand for election himself many times—by all three parties—but he preferred the more powerful background role of expert counselor and adviser Yet he had, in fact, so many of the traits of the politician and they seemed to come so naturally and were so much a part of his personality that it is hard to think that they could not have influenced his advice

He was an opportunist who reacted to events immediately and directly, and his reaction was to produce an answer, to write a memorandum, and to publish at once He was into everything, be it the German mark or the French franc, birth control, the Lancashire cotton trade, buying British, economic sanctions, compulsory savings for a joyful hereafter In the World War II treasury, he nearly drove some of his colleagues crazy with his propensity to keep a finger in every pie "Don't just stand there, do something," would have been his present-day motto. Discussing unemployment in 1930, he said "If we just sit tight there will be still more than a million men unemployed six months or a year hence That is why I feel that a radical policy of some kind is worth trying, even if there are risks about it" (1930*f*) Both Conservative and Labour governments—in the "fatalistic belief that there never can be more employment than there is," as Keynes said (1929*a*)—sat tight over the twenties and thirties, instructed by the civil servants of the treasury school whom he once characterized as "trained by tradition and experience and native skill to every form of intelligent obstruction" (1939)

Keynes instead was ready with ideas—ideas of his own, and the current ideas of others that he made his own—and ready with recipes for trying them They were practical, inventive solutions—such as his proposal to use legislation, originally framed to permit the government to stockpile war materials, for the additional purpose of minimizing price fluctuations. Characteristically, the proposal might be presented as a three-, four-, five-, or more point agenda. He explored alternatives, giving what he considered the preferred order of adoption Often he tied up a package deal; he was always happy if he could pick off two birds with one stone and get in a little social benefit while solving an economic problem He was always ready to commit himself to definite figures—unkind persons said that he made them up Other people say that he was very good at making them up, and that he had a sense, a feel, for what the right figure ought to be. In any case, he cannot be blamed entirely for making up figures, for well into the 1920s there was a great scarcity of economically relevant statistics; it was Keynes who laid the foundations for providing this branch of economic information in Britain.[2]

In doling out his economic prescriptions Keynes's style was to make a direct appeal to action—to governments, to heads of state (as in his "Agenda for President Roosevelt") or to individual citizens. Often it was an appeal for internationally concerted action. At the planning stage of the World Economic Conference of 1933, he urged all governments *simultaneously* to adopt programs of public spending, supported by gold certificates issued by the Bank for International Settlements, to restore world prosperity "What is the charm to awaken the Sleeping Beauty, to scale the mountain of glass without slipping? If every Treasury were to discover in its vaults a large *cache* of gold . would that not work the charm? Why should not that *cache* be devised? We have long printed gold nationally Why should we not print it internationally? No reason in the world, unless our hands are palsied and our wits dull" (1932*c*).

This call to action is typically phrased As Keynes himself said "Words ought to be a little wild—for they are the assault of thoughts upon the unthinking" (1933*b*) Talking on the radio about unemployment in 1931, he hazarded the guess that whenever you saved five shillings you put a man out of work for a day, on the other hand, whenever you bought goods you increased employment "Therefore, Oh patriotic housewives," he paeoned, "sally out tomorrow early into the streets and go to the wonderful sales which are everywhere advertised You will do yourselves good—for never were things so cheap, cheap beyond your dreams Lay in a stock of household linen, of sheets and blankets to satisfy your needs And have the added joy that you are increasing employment, adding to the wealth of the country because you are setting on foot useful activities, bringing a chance and a hope to Lancashire, Yorkshire, and Belfast" (1931*c*)

I have not consulted the contemporary newspapers to learn whether this appeal showed up in the department stores' takings in the January sales, but one result is recorded—the cartoonists had a field day depicting the middle-class "little woman" sallying forth on a spending spree, loading up hubby with parcels and saddling it all on Mr Keynes Were these words a little too wild? In a radio discussion two years later, Keynes was careful to explain the difference between hoarding and *useful* saving (1933*a*).

---

[2] For example, in editing the *Manchester Guardian Commercial* "Reconstruction in Europe" supplements in 1922, Keynes devoted several pages in each issue to following current trade fluctuations, making use of the London School of Economics and Harvard "Business Barometers" In the 1940s the establishment of the Department of Applied Economics at Cambridge owed much to his influence.

But I should give you the rest of the quotation about words. Keynes finished it by adding, "When the seats of power and authority have been attained, there should be no more poetic license."

He liked to call for timely action—*now is the time* to buy sheets, to appoint a board, to settle the world's currency system—and he had his favorite words. He was fond of using the phrases "the prospects for" and "the progress of" and the word "consequences" in his titles. He was always ready to present his "drastic remedy" or his "radical plan" or to approve a "bold measure"—he fancied himself as the *enfant terrible* He was optimistic—where others saw the beginning of a long industrial decline, he felt the country was "in the middle of a painful adjustment" (1930*b*) Addressing international delegates to a meeting of the National Council of Women (his mother was president) in June 1930, he explained that England was suffering "from a sort of *malaise* of wealth"—saving money faster than she was spending it, economizing on the use of labor faster than finding outlets for it, raising the standard of living a little too fast—and he described these phenomena as "the growing pains of progress, not the rheumatism of old age" (1930*c*)

On this occasion, he talked of England's social achievements, in which he took pride He had a strong vein of patriotism, in spite of his internationalism during the twenties He was perpetually fussing about the possibility that British lending abroad was diverting funds from investment in home industry (Was he, subconsciously, like the colonel in "Dr Strangelove," worrying about the sapping of "precious bodily fluids"?) He did not want to see his country become a rentier nation and miss out on the action, he had his own ideas about how Britain should honor her debts before devaluation, and he looked to her to take the initiative and leadership in all his international proposals

He enthusiastically supported any leader whom he thought could make the running After the terrible things he had said about Lloyd George, that goat-footed Welsh Witch, in connection with the Treaty of Versailles, he supported him vigorously in the election of 1929 and wrote the pamphlet that became the textbook of the Liberal campaign (Keynes and Henderson 1929). When taxed with what he had said about Lloyd George's conduct of the coupon election, he replied: "I oppose Mr. Lloyd George when he is wrong, I support him when he is right " It was the same with Winston Churchill. And in December 1930 Keynes commended the enterprising spirit of Sir Oswald Mosley in putting forward a national economic plan (1930*e*), but he did not write about Mosley after that.

He was flexible Having, after much thought, deserted free trade to recommend a revenue tariff with all his might, the moment that England left the gold standard and made such a tariff economically unnecessary, he dashed off a letter to the London *Times* calling attention to the now primary importance of devising a sound international currency system; the discussion of domestic protection should wait until later, he said (1931*a*). And, as the rest of the country breathed a half-sigh of relief after painfully climbing out of the slump, he looked ahead in two *Times* articles to ask how to avoid a future occurrence of such a situation (1937)

Flexibility—or inconsistency? "Inconsistency" was the word that he was branded with When he was attacked for coming over from free trade to a revenue tariff, he lampooned his critics

"I seem to see the elder parrots sitting round and saying 'You can *rely* upon us ' Every day for 30 years, regardless of the weather, we have said, 'What a lovely morning!' But this is a bad bird He says one thing one day and something else the next" (1931*b*)

Even Keynes's critics had to admit that he never tried to pretend when he changed his mind And he usually explained just why he had done so But he was not always so flexible When he was still in his thirties, he was youthfully uncompromising, insisting that the Treaty of Versailles must be revised, unwilling to live with it and accept it as a political necessity As time went on, he became more amenable to compromise "My own view is that I want as much as I can get," he wrote of a League of Nations proposal in 1930, "but I do not want to wreck the whole project by asking for more than I can get" (1930*a*) In the course of time, he even came to tone down—publicly, that is—some of his caustic language Writing to the literary editor of the *Daily Mail* about some small alterations that he wished to make in an article on England's war debts to America, "chiefly with the object of avoiding strong language," he replaced the phrase, "when one reads the rubbish reported from Congressmen, much of it altogether beneath the intelligence and dignity of human nature," by the much milder observation, "when one reads what Congressmen say to reporters" (1932*b*)

Between the two wars he was very visible, energetic, and vigorous on the surface of English political life Articles, pamphlets, and books appeared perfectly timed for the opening of a conference, the preparation of the Budget—or perhaps the preparation of the public for the inevitable economic consequences of some past decision or event The political pamphlets were priced at 6d and 1/-, *The General Theory* sold for 5/- Margot Asquith objected to the publication of the article on the American debts in the *Daily Mail* "You sd have sent it to the *Times*, as those who read the *Daily Mail* are mostly in the Servants Hall (I never take it in)" (Asquith 1933) He was careful to send his articles, punctiliously presented, to the London *Times*, the *Manchester Guardian*, the *Nation*, and the *New Statesman*, at the same time, he never missed an opportunity to publish the same material in more swashbuckling form in the *Daily Mail*, the *Daily Express*, and the *Evening Standard*

But as with every real politician, nine-tenths of the iceberg was invisible Only the small group of those in the know could have been aware of his leadership in 1920 in organizing an appeal to the League of Nations for an international loan and also the first international conference of economic experts in Brussels—civil servants, not politicians; his part in the thinking and research which went into the Yellow Book that reviewed and restated the objects and aspirations of the ailing Liberal party, the questioning that led to his adoption of the revenue tariff policy in the behind-doors Economic Advisory Committee, months before he expressed this change of heart in public, his crucial dominance of the Macmillan Committee of Inquiry into Finance and Industry And just like every practicing politician, he maintained contact with an extensive network of influential friends and acquaintances to whom he could go for help if necessary, because they would do the same with him. His correspondence files are almost a *Who's Who* for an era—not only in the fields of economics and politics, but also in society, literature, and art The day that he surfaced and was observed lunching alone with Ramsay MacDonald, the stock market rose

All these characteristics show Keynes as a natural born and highly efficient, ever alert politician. With his Eton Pop, Cambridge Union background, he was born to the purple, if this was what he wanted Yet he insisted that politics was not his role, that he was more valuable in his chosen capacity as adviser It was a case of emotion versus intellect, some of his contemporaries thought of him as an emotionless, coldly logical machine. Writing of Lloyd George's political craft with introspection into his own personality clearly in mind, he remarked. "A preference for truth or for sincerity *as a method* may be a prejudice based on some aesthetic or personal standard, inconsistent, in politics, with practical good" (1922*c*) In middle age, he was able to write, "I still suffer incurably from attributing an unreal rationality to other people's feelings and behaviour (and doubtless to my own, too)" (1949) His radical approach—"The Republic of my imagination lies on the extreme left of celestial space" (1926)—seemed to destine him for the Labour rather than the Liberal camp, but he shied away from Labour dogmatism and anti-intellectualism To proclaim any dogma as infallible and applicable in all cases, he said, debating the merits of the two parties, was "voluntarily to shut oneself out from any scientific approach to economic problems by means of experiment and investigation" (1928). Also, the Labour party was a class party, Keynes said, "and the class is not my class the *class* war will find me on the side of the educated *bourgeoisie*" (1925).

In considering Keynes, one can never forget his social background. The son of cultivated, thrifty, donnish parents—nonconformist in outlook though apparently not overtly religious—brought up at the close of the nineteenth century in comfortable middle-class Cambridge, polished and finished at Eton and King's, and sophisticated by Bloomsbury and the high-ranking civil service, making and losing a fortune and reestablishing himself again before he was 40—these were personal experiences that made him what he was The man had a confidence, springing partly from his parents' established place in Cambridge society but more powerfully from the consciousness of his own intellect which won him an even more exalted place in the outside world before he was 35.

As the son of his philosopher father and socially conscious mother, he sought in politics a party that in the changing conditions of the time would create a society both economically just and economically efficient, while still preserving individual liberty (1927). He stayed with the Liberals, the party of his parents, by standing intellectually outside of it, despite his campaigning and committee work and participation in summer schools. From his parents he inherited the late Victorian nonconformist belief in the necessity and possibility of the improvement of society by the application of reason and the sense of obligation to one's social inferiors that went with it (The same attitudes appear in the Fabianism of the Webbs, but Keynes seems to have regarded the Webbs as a little naive)

Large-scale unemployment was the basic problem that Keynes came to focus on His received view of the world was of a society in which each man had his appointed place, and it was an injustice for him not to be allowed to feed his family and retain his self-respect in fulfilling the task ordained for him Economic theory then taught the belief that if there were not enough jobs, the laboring force would divide up the supply of work, driving down wages in the process—a hardship to the worker and no challenge to the employer to make himself more efficient, Keynes said His solution, when he wrote "The Question of High Wages" in 1930 (1930*d*), was to increase the workers' real wages by providing social services He told a story about the little girl who, asked if the poor should be made like the rich, replied, "No, it would spoil their characters" (1929*b*) The story was offered tongue-in-cheek, but does one detect the social worker's instincts of his mother?

An achievement that gave Keynes great satisfaction was his wartime scheme for compulsory savings or, as he later chose to call it, "deferred pay " He considered that he had made it "outrageously attractive" to the working class (1940*c*), although the working class seems to have been singularly unenthusiastic about it He regarded it, characteristically, not just as an expedient for financing the war, but as an opportunity to demonstrate the difference between the totalitarian and the free economy. "For if the community's aggregate rate of spending can be regulated, the way in which personal incomes are spent and the means by which demand is satisfied can be safely left free and individual . the only way to avoid the destruction of choice and initiative, whether by consumers or by producers, through the complex tyranny of all-round rationing This is the one kind of compulsion of which the effect is to enlarge liberty. Those who, entangled in the old unserviceable maxims, fail to see this further-reaching objective have not grasped, to speak American, the big idea" (1940*b*).

Toward the end of *How to Pay for the War*, he sums up the scheme as "the perfect opportunity for social action where everyone can be protected by making a certain rule of behavior universal" (1940*a*) It is economically just, it is economically efficient—and the smell is undeniably paternalistic. Keynes grappled with the problems of the twentieth century, but he was born in a big house in Harvey Road, Cambridge, in 1883, and he never really extricated himself from the view of society, and his own position therein, in which his parents had reared him

Unemployment was a problem that according to orthodox theory should not exist, but it was the problem that would not go away He grappled with it with a moral indignation and persistency that conjures up the spirit of another great grappler, John Bunyan, the subject of a biography by Keynes's maternal grandfather, the nonconformist minister, John Brown "Is not the mere existence of general unemployment for any length of time an absurdity, a confession of failure, and a hopeless and inexcusable breakdown of the economic machine?" Keynes demanded in a radio dialogue with Sir Josiah Stamp Stamp, so addressed, observed "Your language is rather violent You would not expect to put an earthquake tidy in a few minutes, would you?" (1930*f*)

It took the years from 1923 to 1936 for Keynes to tidy up the theory of the earthquake In an address on the occasion of the centenary of the death of Malthus, Keynes quoted Malthus himself on the relation of experience to theory, distinguishing between that partial or confined experience that a man gains "from the management of his own little farm, or the details of the workhouse in his neighbourhood"—which is "no foundation whatever for a just theory"—and "that general experience, on which alone a just theory can be founded" Keynes claimed for Malthus "an unusual combination of keeping an open mind to the shifting picture of experience and of constantly applying to its interpretation the principles of formal thought" (1935) Here, in his emphasis on constant referral back to the facts, he stated his own ideal of how a social scientist ought to work

As I have documented earlier in this paper, Keynes behaved from day to day in public like a working politician In private, he was deeply and seriously concerned with the science of economics. He kept coming back to the central problem that the existing theory would not explain At first, he was content to dazzle by demonstrating that a fuller knowledge of economic theory and statistics than his professional colleagues possessed would support different conclusions than their understanding of the economic orthodoxy had led them to But eventually, his intellectual honesty and his concern for economic science brought him to believe that it was not the incompetence of the ordothox economists that was at fault, but the received theory itself. Modifications within the framework of orthodox economics were not enough; a frontal attack on the framework itself was required

So Keynes produced the General Theory as a proof—by the standards of the prevailing economic orthodoxy itself—that, contrary to orthodoxy, the normal state of economic society was not full employment, but general unemployment As a corollary, government policies to raise the level of employment were not mistaken and arbitrary interferences with a well-functioning and efficient economic machine, instead, government interference was absolutely necessary in order for the machine to work at all Thus, Keynes behaved as a scientist in the crucial sense having found the existing body of scientific knowledge in economics increasingly unsatisfactory as a tool for solving the problems that he considered important, he produced a new and rival theory that would explain the discrepancies between the orthodoxy and the facts of observation—discrepancies which formerly had to be explained as special cases.

Recalling Malthus on partial and general experience, how much did Keynes's enthusiasms, his inventiveness, blind his science? One of his early reactions to unemployment was to put it down to overpopulation (1922*a*; 1923), a critic accused him of letting his advocacy of birth control affect his conclusions (*New Statesman* 1923) If he was less of a patriot would he perhaps have been so troubled by the idea of foreign lending? If his upbringing had not inclined him to think of foreign food as less wholesome than home grown, or if he had not bought a Sussex farm for himself and invested a lot of his college's money in a large farm estate in Lincolnshire, would he, after he reverted from his 1931 recommendation of a revenue tariff to his lifelong belief in free trade, have made an exception for tariff protection to agriculture? If his father had been a docker, and not a well-to-do academic philosopher, how would he have dealt with unemployment?

Keynes's origins and place in society strongly influenced and limited his scientific thought There is a passage in *The Economic Consequences of the Peace* where he describes the expectations of a man of his station in pre-1914 society while he worked on his papers, sipping his morning tea in bed, other people cooked and washed and cleaned, providing all his wants—even if he needed money to travel, he could send his servant to the bank (1919) Keynes assumed, extrapolating from his own experience, that "any man of capacity or character exceeding the average" could attain this position.

In the world as it was then, security and independence were the lot of a few; the housewives whom Keynes urged to go out and buy sheets were middle-class housewives with money in the bank to provide for the future Yet, in his ideal society, he desired security and independence for everybody. However, he could conceive of it only in terms of his own experience· social happiness was employment for every one, each in his appointed place, his own niche

Keynes, armed with all the advantages of his upbringing, believed unquestioningly that anyone in England with enough ambition could rise to his proper position in society Nor did he have any question as to whether it was just for a man who had risen as rapidly as he had to be able to rely so implicitly on the full-time dutiful service of others who did not have those advantages (What were the thoughts of the servant, who earned perhaps £1 a week, when he was sent to the bank to fetch £10 or £50? Keynes was fortunate if the servant, like himself, did not bother his head about it) Social injustice existed only in there not being enough jobs to go around If there were servants with talents or character above their appointed stations, they deserved help, financial or advisory, from their betters, it was the obligation of the employer and ultimately the state to make it up to them in welfare benefits and other social transfers

So that, although Keynes thought of himself as a radical, one can see that he took a conservative, even an archaic view of society His "radical" solution of government maternalism has now become the received orthodoxy Even so, his social philosophy made a great leap forward he said that it is wrong for a government to expect people to study, work hard, be honest and responsible, in order to fit themselves for a place in a society presumed to be anxious to employ and make good use of them, when in fact the economic policy of that government precludes a large number of its citizens from having an opportunity for employment and a decent career, government has a responsibility to society to follow an economic policy that will satisfy these expectations

We are left with two problems that did not trouble Keynes We still have not solved the problem of equal opportunity, which Keynes took as a matter of course We do not yet know how to reconcile the boring nature of many jobs with freedom for the human spirit, Keynes, who thought about it in terms that reflected his own social background, looked forward to more automation, less work, and the enrichment of leisure time by cultural activities provided by the state

In summary, in my judgment, Keynes was a politician, but a politician whose constituency was not electoral but intellectual—he had to be a scientist to be a politician And he was a good enough scientist, with a strong enough sense of scientific integrity, and a strong enough aesthetic preference for truth, to recognize eventually that the social science he knew was not good enough to solve the problems he recognized as politically important and that he had to reform the science to make it politically relevant and useful. He was a scientific political economist One can emphasize either the "scientific" or the "political"—and which adjective one emphasizes depends on whether one is writing a political biography of the man himself or a history of economic thought—but both adjectives are appropriate and both are necessary to characterize what the man was and what he contributed to British society and British social history

References

Asquith, Margot. Unpublished letter. Keynes Papers, Marshall Library, Cambridge University, January 2, 1933.

Keynes, J M *The Economic Consequences of the Peace* (1919). In *Collected Writings of John Maynard Keynes* Vol 2. London· Macmillan St Martin's (for Royal Econ Soc), 1971

——— "An Economist's View of Population." In "Reconstruction in Europe " Suppl. 6, *Manchester Guardian Commercial*, August 19, 1922, pp 340–41 (*a*)

———. "Reconstruction in Europe An Introduction " In "Reconstruction in Europe." Suppl 2, *Manchester Guardian Commercial*, May 18, 1922, p 66 (*b*)

——— *A Revision of the Treaty* (1922) In *The Collected Writings*. Vol 3 London: Macmillan St Martin's (for Royal Econ Soc), 1972, also in *Essays in Persuasion*, in *The Collected Writings* Vol 9 London Macmillan St Martin's (for Royal Econ Soc), 1972 (*c*)

——— "Population and Unemployment " *Nation and Athenaeum*, October 6, 1923, pp 9–11

——— "Am I a Liberal?" (1925) In *Essays in Persuasion*, in *The Collected Writings*. Vol 9 London Macmillan St Martin's (for Royal Econ Soc), 1972

——— "Liberalism and Labour" (1926) In *Essays in Persuasion*, in *The Collected Writings* Vol 9 London Macmillan St Martin's (for Royal Econ Soc), 1972

——— "Liberalism and Industry " In *Liberal Points of View*, edited by H L Nathan and H Heathcote Williams London Benn, 1927

——— Report of Liberal summer school debate with Tom Johnston, M P *Manchester Guardian*, August 4, 1928

———. "Mr J M Keynes Examines Mr Lloyd George's Pledge." *Evening Standard*, March 19, 1929 (*a*)

——— "Mr Keynes's Lecture" (letter) *Manchester Guardian*, November 19, 1929 (*b*)

——— "The Draft Convention for Financial Assistance by the League of Nations" (letter) *Nation and Athenaeum*, April 5, 1930, p 11 (*a*)

——— "Investments Abroad" (letter) *Times* (London), March 13, 1930 (*b*)

——— "The Prosperity of England Mr Keynes's Reassuring Picture" (report) *Manchester Guardian*, June 28, 1930 (*c*)

——— "The Question of High Wages " *Polit Q* 1 (January 1930) 110–24 (*d*)

——— "Sir Oswald Mosley's Manifesto " *Nation and Athenaeum*, December 13, 1930, p 367 (*e*)

——— "Unemployment" (broadcast discussion with Sir Josiah Stamp) *Listener*, February 26, 1930, p 383 (*f*)

——— "After the Suspension of Gold" (letter) *Times* (London), September 29, 1931 (*a*)

——— "Economic Notes on Free Trade II A Revenue Tariff and the Cost of Living." *New Statesman and Nation*, April 4, 1931 (*b*)

——— "The Problem of Unemployment" (broadcast talk) *Listener*, January 14, 1931, p 46 (*c*)

——— "The Dilemma of Modern Socialism " *Polit. Q* 3 (April–June 1932) 155–61 (*a*)

——— Unpublished letter to R. J Frew Keynes Papers, Marshall Library, Cambridge University, December 10, 1932 (*b*)

——— "The World Economic Conference, 1933 " *New Statesman and Nation*, December 24, 1932, p. 826 (*c*)

——— "Saving and Spending" (broadcast discussion with Sir Josiah Stamp) *Listener*, January 11, 1933, pp. 41–42. (*a*)

———. "National Self-sufficiency," pt. 5. *New Statesman and Nation*, July 15, 1933. (*b*)

———. "The Commemoration of Thomas Robert Malthus. The Allocutions III. Mr Keynes." *Econ. J* 45 (June 1935) 234 Reprinted as "Robert Malthus. Centenary Allocution " In *Essays in Biography*, in *The Collected Writings*. Vol. 10. London. Macmillan St. Martin's (for Royal Econ. Soc.), 1972.

———. "How to Avoid a Slump " *Times* (London), January 12 and 13, 1937

———. "Democracy and Efficiency" (interview with Kingsley Martin). *New Statesman and Nation*, January 28, 1939, p 122

——— *How to Pay for the War A Radical Plan for the Chancellor of the Exchequer* London Macmillan, 1940. Reprinted in *Essays in Persuasion*, in *The Collected Writings* Vol. 9. London Macmillan St Martin's (for Royal Econ. Soc.), 1972. (*a*)

——— "Mr Keynes's Plan Control of Boom and Slump" (letter) *Times* (London), April 10, 1940 (*b*)

———. Report of talk to the Fabian Society Keynes Papers, Marshall Library, Cambridge University, February 21, 1940. (*c*)

——— "My Early Beliefs " In *Two Memoirs*, with introduction by David Garnett London. Hart-Davis, 1949 (posthumous) Reprinted in *Essays in Biography*, in *The Collected Writings* Vol 10. London· Macmillan St Martin's (for Royal Econ. Soc), 1972.

Keynes, J. M , and Henderson, H D *Can Lloyd George Do It? An Examination of the Liberal Pledge* London *Nation and Athenaeum*, 1929 Reprinted in *Essays in Persuasion*, in *The Collected Writings* Vol 9 London Macmillan St Martin's (for Royal Econ Soc.), 1972.

New Statesman "Malthusian Moonshine" (unsigned article) September 22, 1923, pp 664–65

Group Segregation and Optimal Jurisdictions

Martin McGuire

University of Maryland

Even if people are entirely devoid of any feelings toward each other (sympathetic or antipathetic) at a personal level, they may find it in their interests to set up and join associations which are segregated on the basis of income and/or tastes This paper constructs a model to show how wealth, technology, and preferences can interact to provide common incentives for people to form segregated groups for the provision of local collective goods—that is, goods which cannot readily be supplied and priced on a variable unit-of-services basis, but which are best provided communally at the same level to all members of the association, in such a way that nonmembers are excluded from enjoying them altogether

I. Introduction

The study of groups is of particular interest to economists because economics has something special to add to our understanding of group formation Our population divides itself into groups in a variety of ways, according to color, language, ethnic origin, age, marital status, etc , in many cases such stratification may involve no individual choice When groups arise as voluntary associations, however, for the mutual benefit of the membership, analysis of the basis for and the structure of such mutual cooperation is essentially economic Economic reasoning about groups will suggest two alternative "pure" paradigms to explain segregation. In both these pure cases, people cooperate for their mutual benefit, the alternatives differ with respect to the source of those mutual benefits

Earlier versions of this paper were read to the Committee on Urban Public Economics, where I benefited from members' criticisms and to the Barcelona meetings of the Econometric Society in August 1971 I also want to thank Thomas Schelling, Henry Aaron, Hirofumi Shibata, and the referees for comments. Assistance from the Bureau of Business and Economic Research, University of Maryland, from the Institute of Social and Economic Research, University of York, England; and from the National Science Foundation is gratefully acknowledged.

On the one hand, reciprocal benefits from association may stem solely from the gratification of personal tastes for association—the pure pleasure of being near others one wants to be near.[1] A second source of mutual benefit from voluntary association, and, therefore, a second pure case or paradigm for group formation, is the common advantage people may find in producing, exchanging, or consuming some good which they all value In this second pure case, personality is nothing; no benefits derive from association as such; people confer no externalities upon one another solely by virtue of their existence Doubtless both kinds of segregation often reinforce each other in the real world, but it is useful to isolate the nonpersonal mechanisms of group segregation for closer analysis, so as to appreciate the power of people's concern with the utility of things to isolate them from other people

The purpose of this paper is to build such an apersonal economic model of group formation based on tastes, wealth, and technology More specifically, we explore the logic of group segregation and the interaction among these determinants in the provision of *local collective goods or services*

II. Optimum Jurisdictions, Public Good Supply, and Taxes for a Homogeneous Population

Historically, the greater part of economic analysis has approached the theory of local collectives from two perspectives first from the viewpoint of pure public good theory, questions of optimal supply and tax distribution have been analyzed (McGuire and Aaron 1969, Samuelson 1969) Second, the effects of interjurisdictional externalities upon public good supply and burden-sharing optima have received considerable attention (Pauly 1970*a*, Williams 1966) Less attention, however, has been given to the question of the size and composition of each consuming group and to the interaction of size and composition with optimal supply

One reason for the relative neglect of this question in the literature is that it is harder to analyze As an example of this, take the problem of defining an economic optimum In the case of pure public goods, the size and nature of the consuming group is given in advance, and the problem is to determine how much to supply, how to distribute cost burdens, and how, if at all, to use prices and quasi markets to attain desired outcomes. On the other hand, to define an economic optimum in the provision of a local public good, several problems must be resolved simultaneously· (1) the optimal number and shape of jurisdictions; (2) the best way of distributing households among jurisdictions; and (3) the resulting patterns among local collectives: to include size, tax rate, and public good supply distributions, and homogeneity or heterogeneity of local population groups. This paper provides a unified theory of such simultaneous optimization, thereby consolidating extant models by Tiebout (1956, 1961), Buchanan (1965), and Rothenberg (1970) [2]

---

[1] Thomas C Schelling (1969) has explored this basis for group formation with a set of pure personal-taste models, models in which the aggregate outcome of individualistic behavior follows from the nature, the diversity, and the distribution or structure of personal tastes for association

For our purposes, a local public good is one which can be produced in various quality or quantity levels,[3] but must be provided uniformly[4] to all individuals consuming at the particular level in question An individual is assumed to choose only one quality or quantity level for each type of public good and thereby to identify himself with the particular level chosen Examples of such situations are very common police protection is supplied to all citizens within a county (town, precinct) approximately uniformly, so are fire protection within a fire district and water and sewer service within a sanitary district, in most places equal access to one level of primary and secondary education is available to all within a particular school district[5]

---

[2] Writing in 1956 and again in 1961, Tiebout proceeded on the assumption of a mobile "consumer" and fluid, redefinable local jurisdictions which effect no externalities upon each other His conclusions were

1 For any configuration of jurisdictions (group size, collective output, tax rates, and expenditures) each consumer-voter "moves to that community whose local government best satisfies his set of preferences The greater the number of communities and the greater the variance among them, the closer the consumer will come to fully realizing his preference position" (1956, p 418)
2 For every pattern of community services, there is an optimal (that is, average cost minimizing) community size Optimal sizing of communities will result from the combined motives (*a*) of individuals to leave higher than minimum average-cost communities, (*b*) of city managers to restrict entry through zoning, collusion with realtors, etc
3 People with similar fiscal tastes given the costs to each of them should move together

More recently, James Buchanan (1965) analyzed a similar if not identical set of problems in "An Economic Theory of Clubs" (the formal similarity between this and Tiebout's model has not been widely recognized in the literature—possibly due to the fact that Buchanan's analysis was not developed exclusively in the context of local government organization) Although Buchanan explicitly assumed that the various members of a collective impose external economies or diseconomies upon each others' *consumption*, he might alternatively have followed Tiebout and have treated externalities as arising in the *production* of the collective good In any case, Buchanan's central conclusion closely parallels Tiebout's, namely, that for a homogeneous group of individuals the optimum size of jurisdiction (or club) and level of collective good must be simultaneously determined, and that freedom of association—freedom both to create a club and to join one already there—will lead utility maximizers to just such an optimum More recently still, Rothenberg (1970) discusses the local collective problem as one of reaching a balance between consensus among members of particular jurisdictions and scale economies in production Jurisdictional consensus or homogeneity requires small, inefficient groups, he conjectures, while "efficiency" entails large, heterogeneous, conflict-ridden units

[3] The "amount" of the good will be taken as a continuous variable in this study representing either *quantity or quality* of the good or service

[4] Or in fixed proportions! Our analysis will assume equal provision of the good to all members in the "local" group, although the analysis could easily be extended to include diversity in the fixed proportions in which the good is distributed. So long as the distribution of a good can, in fact, be regulated according to a fixed proportion rule, the good can be treated as "public" (see Buchanan 1968, esp pp 50–56, entitled "Private Goods as Public Goods")

It will prove useful to begin our analysis with a case in which segregation is impossible or meaningless, that is, the case of a homogeneous national population with identical tastes, incomes, and, therefore, demand curves for local public goods Consider a world of two goods, y, the private-good numeraire and x, the local public good Everyone's utility function can be written as $U^i(x^i, y^i)$, $i = 1 \cdot \cdot P$, P = population size

Next we assume that a certain cost function for a given jurisdiction applies universally to all possible jurisdictions whatever their composition[6]

$$C = C(x, n),$$

where C = total cost measured in units of the numeraire, x = level of public good supplied to each member of the jurisdiction, and n = population or size of jurisdiction Writing the total cost function in this way suggests that our concept of "jurisdiction" is closely analogous to the idea of a firm in the conventional analysis of private goods In our case, however, individuals are identified with their particular "firm" They cannot (or merely do not because of high cost) divide their consumption between jurisdictions Let us assume that every member of this collective has the same opportunity to enjoy the public good x People at the fringes of the jurisdiction do not benefit less in any way than do individuals at its center, people outside the jurisdiction receive no benefit from x.[7] Thus, in our model the essence of the "localness" of the "public" good is captured by making total cost a function of n, the subpopulation served by (or size of) the jurisdiction[8] The influence of n upon C or upon x may be regarded either as purely a technical production effect or alternatively as a congestion-consumption effect

Clearly, we must concern ourselves with the case where the total cost of public-good production increases with size of the jurisdiction (as well as with the level of public good), or, equivalently, the quantity or quality of public good supplied is degraded as numbers served increase. That is,

---

[5] These examples are all of publicly provided goods There is a close analogy between these public good "jurisdictions" and a private good market area The analogy is discussed in more detail in McGuire (1972)

[6] We are measuring the local collective good, therefore, in "consumption units," each unit of which will figure identically in the utility function of every member of the local collective This is one particular way of representing a general implicit relation $F(C, x, n)$ among the variables cost, output, and group size We might equally have written $x = x(C, n)$ It will become clear that this formulation would produce identical formal results but with a slightly modified interpretation (see n. 12).

[7] We rule out interjurisdictional spillovers in this analysis.

[8] If one allowed for the fact that costs may also depend on the type of user, the conclusions of our analysis might change.

$$\left.\frac{dx}{dn}\right|_{C = \text{constant}} < 0; \qquad C_x \equiv \frac{\partial C}{\partial x} > 0; \qquad C_n \equiv \frac{\partial C}{\partial n} > 0.$$

Without this stipulation our problem would disappear. If numbers served could be increased at zero cost with no quantity-quality loss, one national jurisdiction would be optimal, and the questions of supply and cost shares have been thoroughly treated already for this case of a national public good. As we shall see presently, even with $C_n > 0$, a single national jurisdiction may still be the best way of organizing for collective supply

To proceed with the analysis, let us define c^i as the ith person's tax or contribution to defraying the costs in "his" jurisdiction so that a person's final consumption of the private good is his initial endowment (or earnings) $\bar{y}^i$ less this tax, that is, $y^i = \bar{y}^i - c^i$. We will let $\bar{y}^i$ be the same for everyone. Clearly, optimization of local public-good supply is a meaningful question only with reference to some social welfare objective Therefore, we assume that each of the identical individuals' utilities is equally weighted in a social welfare function This formulation implies a symmetry in the solution to the welfare maximization problem—namely, equal utility levels and public-good consumption for each individual We can exploit this symmetry and simplify the derivation of necessary conditions for a welfare maximum by observing first, that cost burdens must be equal, $c^i = C(x, n)/n$, and second, that maximization of social welfare becomes equivalent to maximizing the utility of the representative individual·

$$\text{Maximize: } U\left\{x, \left[\bar{y} - \frac{C(x, n)}{n}\right]\right\}.^{9}$$

Necessary conditions for a maximum are:

$$nC_n = C: \qquad \frac{C}{n} = C_n: \qquad \text{and} \qquad n\frac{U_x^i}{U_y^i} = C_x.$$

From these two equations in two unknowns we solve for x and n. If the total population is of size P, there should be P/n identical jurisdictions. The first of these conditions requires that each of the P/n groups be sized such that average cost per person C/n be at a minimum; this occurs when average and marginal costs (of adding one more person) per person

---

[9] This illustrates that people simultaneously create external economies and diseconomies for each other in local groups. An increase in n lowers U (external diseconomy), since $C_n > 0$. An increase in n raises U (external economy) since it spreads the cost. In our model people do not have a pure distaste for association—an interpretation which the less specific formulation such as $U = U(x, n, y)$ would allow. Our conclusions, however, apply equally to the less specific formulation.

are equal [10] Hence, each individual's share of the cost burden, c^i, equals the marginal cost he imposes on his local system, C_n

The second of the above conditions is the familiar one that

$$\sum \text{MRS}_x = \text{MC}_x$$

for each and every of the P/n groups (where MRS_x indicates an individual's marginal rate of substitution between goods x and y, that is, U_x/U_y, and MC_x indicates the marginal cost of supplying everyone in the group with one more unit of x)

These equations specify the relationships which must obtain among the variables (1) size of each jurisdiction, (2) level of public good supply, and (3) tax burden The equation

$$C(x, n) - nC_n(x, n) = 0$$

may be thought of as a *normative* supply function, relating size of jurisdiction to level of public-good supply Similarly, let us tentatively interpret equation

$$n \frac{U_x\left\{x, \left[\bar{y} - \dfrac{C(x, n)}{n}\right]\right\}}{U_y\left\{x, \left[\bar{y} - \dfrac{C(x, n)}{n}\right]\right\}} - C_x(x, n) = 0,$$

as a *normative* demand function. This equation indicates how level of public goods ought to vary as the size of jurisdiction increases, if the efficiency condition $\text{MC}_x = \sum \text{MRS}_x$ is to be maintained The first of the above set of equations is called a "supply function" on the grounds that it is purely technological, representing the values of jurisdiction size which minimize costs of supplying various levels of public good uniformly to the homogeneous population The second equation, on the other hand, has been labeled a "demand" equation, since the utility functions of individuals in the homogeneous population figure directly in their derivation

With this information at hand, let us portray the solution to this problem graphically Figure 1*a* gives an illustration of the basic, underlying cost curves $C(x, n)$, and the various points on the supply curve where $C/n = C_n$ In the case illustrated, along that supply function $dx/dn < 0$, hence, as shown in figure 1*a*, $x^0 < x^1 < x^2$, $n^0 > n^1 > n^2$ This progression of per person cost minima means that big jurisdictions are

---

[10] If the population (of size P) cannot be divided into an even number of groups of size n (as seems likely), the optimum-size group will either be the next lowest value of n (to round up P/n to an integer) or the next highest value of n (to round P/n down to an integer) Continuous methods of analysis will not solve this problem (see Pauly 1970).

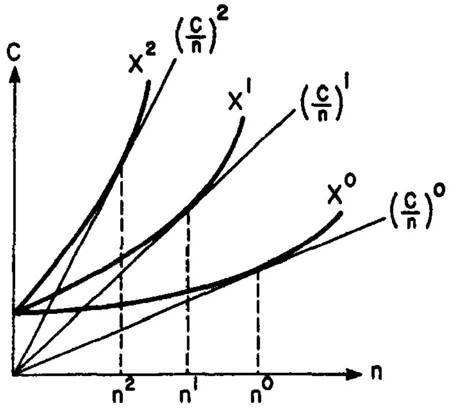


FIG 1*a*

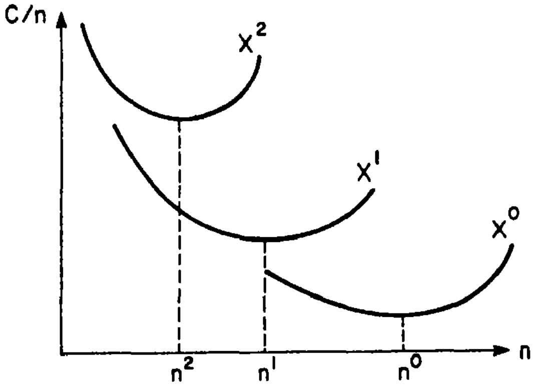


FIG 1*b*

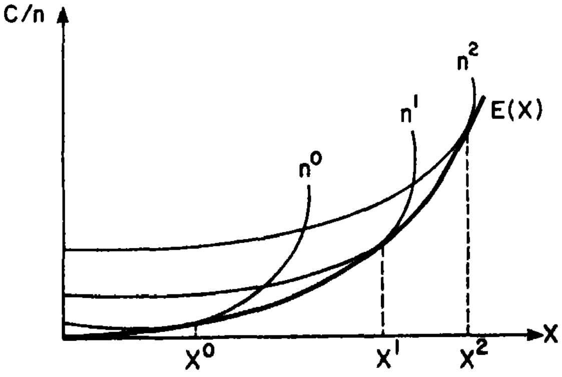


FIG 1*c*

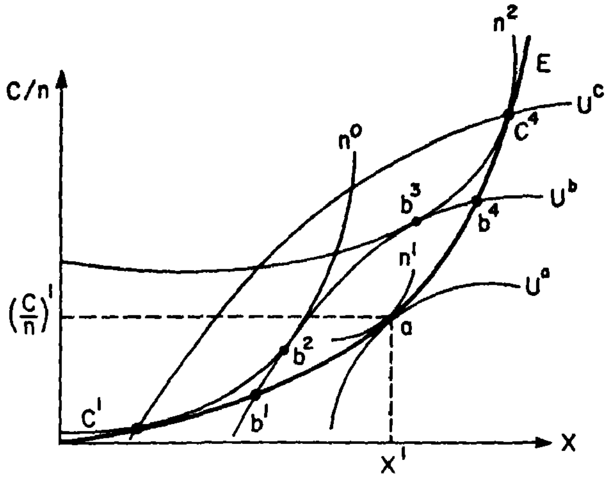


Fig 2

assumed to supply lower levels of public good than do small jurisdictions [11]

Figures 1*b* and 1*c* now show per person costs as a function both of jurisdiction size and level of output, respectively In figure 1*b* we show U-shaped costs per person If the function C/n rose monotonically, the good x should be supplied as a private good, if C/n steadily declined, there should be only one "national" jurisdiction Figure 1*c* shows the set of efficient supply points along an envelope $E(x)$ Each efficient combination (x^0, n^0), (x^1, n^1), etc —satisfying $C_n = C/n$—contributes one point to this technically efficient supply schedule

We can now employ figure 2 to give a complete description of the determination of the optimum jurisdiction size (number of people), optimum level of local public good, and corresponding payment structure for a *perfectly homogeneous population* Several utility contours between income (i e , the negative of C/n) and x are shown for one representative individual in the homogeneous population This individual, and therefore the entire population, maximizes utility by (1) organizing itself into groups of size $n = n^1$, (2) supplying each group and therefore each member with $x = x^1$ of local public good, and (3) charging each member $c^1 = C(x^1, n^1)/n^1$

---

[11] This is an arbitrary assumption Total differentiation of the supply function, $C - nC_n$, leads to an expression for the slope of that function, namely,

$$\left.\frac{dx}{dn}\right|_{\text{supply}} = \frac{nC_{nn}}{C_x - nC_{nx}}$$

which may be positive or negative Our assumption that this slope is negative is made solely for illustration.

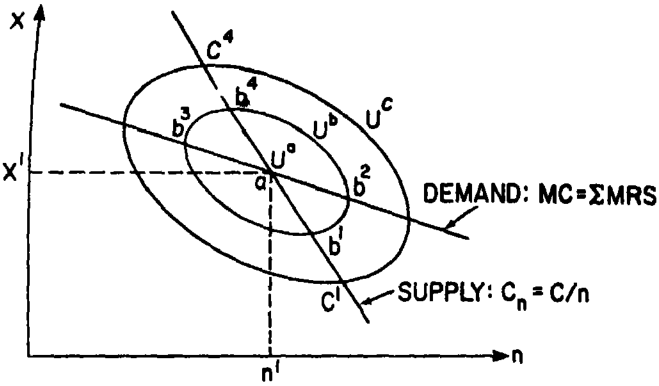


FIG 3

Another useful way of illustrating this optimum is shown in figure 3 Consider one (nonmaximum) indifference curve in figure 2 Let us hold the utility level constant at, say U^b, and plot, in figure 3 the combinations of x and n which just allow U^b to be maintained Clearly, two points on the supply-efficiency envelope allow U^b to be maintained, namely, b^1 and b^4 Similarly, there are two points which allow U^b to coexist with efficiency on the demand side, namely, b^2 and b^3 To see this, suppose the size of jurisdiction is arbitrarily set at $n = n^o$, given that size and assuming n^o identical individuals actually "inhabit" the jurisdiction, the utility-maximizing level of collective output must satisfy

$$\mathrm{MC}_x = \sum_{i=1}^{n^o} \mathrm{MRS}_x^i$$

This output level is identified by the tangency between $C(x, n^o)/n^o$ and the representative individual's utility curve at point b^2

All combinations of x and n which just allow U^b to be maintained form a closed loop in figure 3 [12] At points b^1 and b^4 this loop has a horizontal slope, while at points b^2 and b^3 its slope is vertical This follows from the

---

[12] For every combination of x and C/n two values of n exist except on the efficiency envelope itself, where one value of n obtains This construction effectively shows the unity between Tiebout's and Buchanan's analysis (see n 2), since we have taken Tiebout's assumptions and, in figure 3, produced Buchanan's geometry More generally, this result shows that our analysis applies to any "optimum club-size" problem where the size of membership simultaneously exerts an external economy and diseconomy on each member The fact that we have formally ascribed these simultaneous externalities to the *production* process is more or less incidental Since we end up with a general, representative utility function $U(x, n, y)$ in which the precise mechanisms of externality are not explicit, we could have generated the same general model with some other externality mechanism For example, we could have started off the analysis by defining the local public good as $x = x(C, n)$ thereby implying that the external diseconomy occurs in a "consumption process"

fact that at b^1 and b^4 utility is maximized with respect to jurisdiction size, while at b^2 and b^3 utility is maximized with respect to level of collective good supply. Figure 3 shows several contours of constant utility constructed in a similar way. Now, consider all points along the line b^2b^3. These indicate the utility maximizing level of public-good production for various values of group size. Clearly, the line b^2b^3 represents the efficient demand equation, $MC_x = \sum MRS_x$. Note that this "demand" equation depends on income, on tastes, and on technology.[13]

Next consider all points along the line b^1b^4: these points indicate the utility-maximizing value of jurisdiction size for various values of public good output Thus, the line b^1b^4 represents the efficient supply equation $nC_n = C$ Note that this supply schedule is purely technical Finally, figure 3 shows the simultaneous optimization of jurisdiction size and level of public good supply at the intersection of the supply and demand curves, and utility index U^a.

III. Efficient Jurisdictions for Heterogeneous Populations

We can now turn to mixed populations and to the normative question of whether integration is more or less efficient than segregation when grouping people for the provision of local public goods [14] As an initial simplification of the problem of finding efficient jurisdictions for a heterogeneous population, let us suppose:

1. Population heterogeneity takes the form of multiple subpopulations each homogeneous unto itself and discretely separated[15] from every other homogeneous subpopulation. Every member of a homogeneous subpopulation has the same demand curve $D(x, n)$.

---

[13] Given the technology, this demand curve as drawn in figure 3 will have a shallow (steep) slope if the public good (x) and private good (y) are complements (substitutes). And for given underlying preferences, the "demand" curve will be steep or shallow according as marginal costs C_x increase rapidly or gently (with respect to x)

[14] The foregoing analytics suggest an operational definition of "diversity" or "heterogeneity" Consider the demand curve of figure 3 again. One demand curve can be taken to represent the combined income and preferences of one individual, without regard to the existence of other identical individuals. Clearly, different individuals with different tastes and incomes can have identical demand functions, provided the differences in tastes and incomes cancel. Thus, one definition of "economic homogeneity" may be "identical demand curves." But even if Mr. i and Mr. j have different demand curves, the two could be considered "economically homogeneous" if each's demand curve, intersects the supply curve at the same point. Since at that point both parties agree on (1) size of jurisdiction (2) equal cost shares, and (3) level of public good supply, it is plausible to define the two as homogeneous. [Whether this definition is accepted or not, the underlying argument remains valid that individuals of highly diverse needs, tastes, and incomes may find themselves in unanimous agreement on precisely the three variables (C/n, n, x) which must be simultaneously optimized for jurisdictional efficiency.]

[15] This discrete separation is not necessarily spatial; it is mathematical.

2. Each homogeneous subpopulation is large enough to form many efficient jurisdictions. Hence, we can ignore the problem of where to put "leftover" individuals as being of second-order importance
3. A separate jurisdiction can be formed for each local public good (separate jurisdictions need not be formed if cost and demand conditions are the same for different public goods).

Given these assumptions, the foregoing analysis of homogeneous populations implies that integration—in the sense of forming mixed tax-expenditure jurisdictions of people of diverse demands—is economically inefficient Given the foregoing assumptions, whatever the source of differences in demand by different representative individuals (whether due to diverse fundamental preference structures or to differences in income levels), *it is Pareto-optimal to allow homogeneous (with respect to marginal rates of substitution) groups to form their own collectives, it is nonoptimal to effect income transfers by integrating diverse groups and then varying cost burdens according to equity judgments* [16]

To demonstrate this conclusion, let us first establish the normative proposition that if we integrate diverse classes and simultaneously require that all members of a jurisdiction pay equal cost shares, *everyone* is worse off than he would be if allowed to join segregated groups We can demonstrate this with figure 4 Suppose the population is made up of two distinct classes, class 1 and class 2, of equal size [17] If allowed to form separate, segregated collectives, the first class should organize itself into jurisdictions of size n^1 and supply each member with x^1 in public goods at a cost of C^1/n^1 to each member Class 2 should[18] choose $n^2 < n^1$, $x^2 > x^1$, and $C^2/n^2 > C^1/n^1$, respectively [19] (The income value of x^1 to each member of class 1 is shown as distance $[(C^1/n^1) + a]$, and the income value of x^2 to each member of class 2 is $[(C^2/n^2) + b]$)

Now let the two classes be integrated and required to share costs equally A new intermediate facility size, n^3, should be chosen, supplying all

---

[16] I conjecture that this theorem is also valid for the case of populations with continuous distributions of income tastes As pointed out to me by Joanne Linnerooth, one exception to this general rule exists suppose the income transfers which are judged to be socially desirable just happen to result in an income distribution such that diverse individuals demand the same quantity of public good *at an equal cost share* for this particular case it is just as efficient to "force" integration and effect the desired income transfers through discriminatory cost sharing as it is to redistribute income directly and allow uniform homogeneous jurisdictions to arise "naturally "

[17] To determine the new "integrated" solution we need to know the relative sizes of class 1 and 2 For simplicity we assume the two subpopulations have equal numbers.

[18] Note that although we assume the two classes to be of equal size, each should organize itself (if segregated from the other class) into groups of different size for purposes of public good supply.

[19] For illustration we continue to assume jurisdiction size to be inversely correlated with level of public good supply.

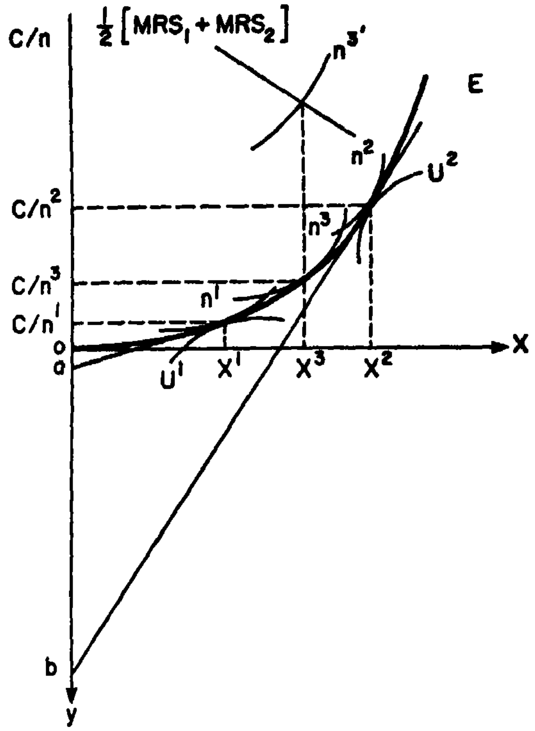


FIG. 4

members of both classes with x^3.[20] If individuals from both classes, once integrated, pay equal cost shares C^3/n^3, each member of both classes will be worse off in the integrated situation than in the segregated one, as the diagram clearly shows. Note that this universal deterioration does *not* stem from inefficient supply in the integrated groups. In each integrated jurisdiction the efficiency conditions $C_n = C/n$, and $\Sigma MRS_x = C_x$ are maintained. Hence, the joint utility loss stems solely from class integration.

It may seem that the inefficiency of integration follows only because we impose the severe restriction of equal cost shares. But this is not so. Differential cost sharing in public good provision is identical to direct income transfers. Consequently, direct transfers of income before integration, and transfers via cost share distribution after integration, in

---

[20] This new solution point is determined by the intersection of the new average MRS curve (demand for good x) with the marginal cost curve $E'(x)$ [marginal to the efficient supply envelope $E(x)$]. The $E'(x)$ is calculated as

$$\left.\frac{d(C/n)}{dx}\right| \text{ given that } C_n = C/n: \qquad \text{hence, } E'(x) = \frac{\partial C/\partial x}{n}.$$

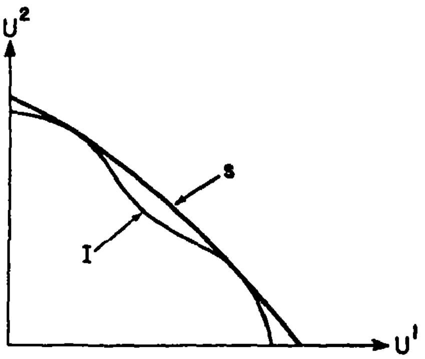


Fig 5

effect are identical redistribution instruments [21] Since we have identified an "efficient" postintegration distribution at which all parties lose (in comparison with the preintegration distribution), there is clearly no redistribution from this inferior point which makes *both* classes better off Furthermore, this argument is valid for every possible preintegration distribution of income, since our choice of such an initial point was totally arbitrary Therefore, the set of possible utility distributions among segregated groups dominates (weakly) the set of such distributions among integrated groups For each of the latter possibilities, a utility distribution among segregated groups exists which makes everyone better off (or no worse off) This conclusion is illustrated in figure 5 for the two homogeneous classes 1 and 2 Assuming we always equalize income among individuals within each class, no matter what interclass distribution is desired, we need only plot the utility (U^1 and U^2) of a representative individual from each class The set of utility possibilities under a segregated regime is shown by the curve S, the set under an integrated and tax-discriminatory regime is shown by I There may be one or more distributions for which the two regimes are equivalent (distributions under which everyone would freely integrate and pay equal shares), but in general S is superior to I

Thus, given our limiting assumptions, we are led to the conclusion that for the provision of local public goods, class isolation is overall efficient, it is more efficient to transfer income directly between classes than to do so indirectly by manipulating jurisdiction membership and cost shares paid for public goods (Of course, income transfers will influence the ultimate configuration of jurisdictions) By the nature of the problem, various jurisdictions cannot export their public goods to others, hence, equalization of marginal evaluations cannot come about via regional or local trade To give a simple intuitive explanation of this conclusion, consider the case in which the optimal membership of a local collective is independent of the level of public good supplied Obviously in this event segregating individuals with identical preferences is superior economically to integrating them We have extended this conclusion to the case in which group size and public good supply interact Evidently the argument is generalizable beyond two classes.

---

[21] For a given membership size within each local jurisdiction, the good provided is a public good Recently it has been shown for this case that manipulation of the level of public goods supply together with the distribution of the costs of supply among members of the community, constitutes a perfectly efficient means of redistributing income—that is just as ideally desirable as lump sum transfers (see McGuire and Aaron 1969)

We may also draw one positive conclusion from an otherwise normative analysis—namely, that when in fact costs of local public good provision are borne equally regardless of wealth, all classes have *symmetrical* incentives to isolate themselves from each other, and migration between groups (even ignoring relocation expenses) will not voluntarily occur (if cost burdens are equally borne), since both groups lose from the population transfer Indeed, since property and income taxes obviously are not levied equally in the real world, we get a partial explanation of why devices such as zoning are used to keep poor people out of rich communities and why (whatever the distributional consequences) zoning for this purpose may be efficient

IV. Diversity of Local Public Goods and Processes of Group Formation

Let us now turn to a process by which—with "freedom of association"[22]—individuals of various preference structures and incomes might form themselves into groups of diverse sizes supplying diverse quantities and qualities of public good A truly adequate positive theory of such processes should have not only to explain grouping patterns, but also differential burden-sharing outcomes as between different individuals within each local collective, since equal taxation of various individuals in the same local collectives is rarely even approximated Unfortunately, such a synthesis of differential taxation and group formation has not yet been obtained In this section, we examine instead a process of group formation on the assumption of equal burden sharing within each group To the extent that tax differentials between groups far exceed those among groups, this exercise will have empirical application, and in any case, the methodological techniques of analysis developed for our more limited process should find use in the analysis of more complex local collective and tax-sharing situations

---

[22] The reader is reminded that individual preferences range only over goods and not over other people per se or association with them Thus, we are explicitly ignoring realistic situations in which "freedom of association" is ambiguous—that is, situations in which one group wants to associate with a second but the second wants to disassociate itself from the first

We take it as axiomatic that an individual choosing a community will compare alternative cost-benefit packages and choose the one with the greatest net gain "In the short run," where the alternative public facilities available are arbitrarily limited in number, entry fees, user fees, or taxes levied as a rationing device may result in some (low-income) groups being totally excluded from access to particular types of facilities For example, consider two populations and two corresponding public facilities If one facility were reserved for each community and suboptimizing user fees were charged, each group would enjoy some access to the public service in question But this outcome might be inefficient, one group might profitably bribe the other to turn over its facility In this latter case a single price would prevail at both facilities, a price which might easily exclude one group from use of any service altogether Following the line of thought suggests that in the interests of equity it may often be justifiable to reserve some facilities for certain interest groups, to charge discriminatory prices among groups for use of the same facility, or to deliberately allow congestion itself (rather than a price-tax mechanism) to regulate total demand

But such limitations on overall efficiency as these are unjustifiable in the long run when group formation and public facility provision are both variables In the long run, *if* the total society is big and varied, an individual making the decision of which local collectives to join (and therefore of his tax and public good consumption level) will be confronted with a wide range of choices, each slightly different from the next As one basic assumption for the analysis of these choices, we take it that each local group is willing to accept new members who pay the full additional cost they would impose on the group,[23] in other words, the price of admission to group k is assumed to be $C_n(x^k, n^k)$ Now in comparing any two groups, an individual will observe that the differential cost between groups k and j is $C_n(x^k, n^k) - C_n(x^j, n^j)$, and the differential benefit is $x^k - x^j = \Delta x$ Assuming the two groups to be fairly "close," the differential cost is approximated by

$$\Delta C = C_{nx}\Delta x + C_{nn}\Delta n$$

Hence, any one individual who is "shopping" for a collective to join will calculate a marginal cost-benefit ratio of

$$\frac{dC}{dx} = C_{nx} + C_{nn}\left.\frac{dn}{dx}\right|_{\text{supply}},$$

---

[23] This full additional cost is the amount of money a newcomer must contribute to insure that his entry into the group leaves the established members' benefits and costs unaffected

and will proceed in his shopping until he finds the group at which this marginal rate of technical substitution equals his subjective marginal rate, that is,

$$\frac{U_x}{U_y} = C_{nx} + C_{nn} \left.\frac{dn}{dx}\right|_{\text{supply}} .$$

As a second basic assumption to carry the analysis forward, we suppose that every member of the population (including old members of jurisdictions who mutually impose opportunity costs on one another) has the option of shopping around among about the the same range of packages and is required to pay opportunity cost, C_n, for his public services[24] Given this further assumption, it follows that the entire (heterogeneous) population will be led to a constellation of diverse groups, each (roughly) homogeneous Each group will tend to contain individuals with the same subjective rate of substitution between money and the public good This clearly is a process of self-selection or segregation by taste and income level Each group will be characterized by an equation,

$$n\frac{U_x}{U_y} = nC_{nx} + nC_{nn} \left.\frac{dn}{dx}\right|_{\text{supply}}$$

Last, let us further assume that each alternative jurisdictional option or tax-expenditure package is efficiently sized—such that in every jurisdiction $C/n = C_n$[25] This assumption, together with free, utility-maximizing search and marginal cost "membership charges," implies an overall-efficient constellation of segregated jurisdictions For under these assumptions, when a single individual compares benefits and costs of alternative jurisdictions, he will observe that size and local service are correlated as (see n 11)·

$$\left.\frac{dn}{dx}\right|_{\text{supply}} = \frac{C_x - nC_{nx}}{nC_{nn}}$$

Substituting this expression into the individual consumer's cost-benefit equation, one obtains

$$n\frac{U_x}{U_y} = nC_{nx} + nC_{nn}\left(\frac{C_x - nC_{nx}}{nC_{nn}}\right) = C_x \quad \sum \text{MRS}_x = \text{MC}_x.$$

Consequently, the process of individual search for the right cost-benefit package will lead to efficient group formation on the demand side In short, given the above assumptions (including equal taxation within groups) we can build a self-sustaining process whereby an initial configuration of efficient jurisdictions is maintained But are these assumptions realistic? Or do individual incentives exist to violate the

---

[24] Again, we ignore relocation costs

[25] As an assumption this has appeal in the context of equal (opportunity cost) taxes for all members of the group. Unless the group is sized efficiently, total collections, nC_n, will not equal total costs, and the group will show a deficit or surplus.

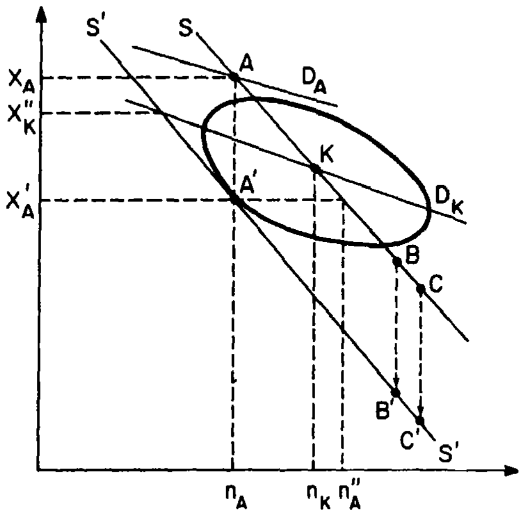


FIG 6

assumptions? And what if we start from an inefficient configuration, would the process of individual choice we describe correct or reinforce such inefficiency?

One incentive tending toward inefficiency is well known, namely, the "free-rider" potential and the resulting tendency for collectives to underproduce public goods [26] We can use figure 6 to illustrate the effect of systematic underproduction of public goods by old residents upon the selection-of-jurisdiction process by newcomers In doing this we also will illustrate one approach to the problem of the interaction between the individual choice to join or quit a local group and the collective choice of public good supply Suppose that an "ideal" distribution of jurisdictions is given by the line *SS* Jurisdictions of type *A* would ideally be inhabited by n_A individuals with demands D_A and would supply x_A and similarly for *B*, *C*, etc Suppose all *A*-type, *B*-type, etc residents, in attempting to exploit the free-rider potential, reduce the public good supplied to x'_A, x'_B, etc Therefore, a new potential member, Mr K, sees a set of possibilities *S'S'* Mr K evidently will choose group *A'*, for in doing so he maximizes his utility Note that Mr K will (on the assumption of equal cost shares).

---

[26] This tendency on the part of established members can coexist with rational selection of benefit-cost packages by newcomers *Here one should emphasize the important distinction between the collective decision of the membership as to the tax-expenditure package and the individual decision of any one person as to which group to join* A local collective good, although "public" to members within the group, is "private" as among groups Thus, a system of local collectives is created and sustained by economic forces stemming from two very different types of decision process As observed recently by Ellickson (1971), a rigorous theory of local public goods must incorporate the interaction between these two types of choice.

1 Join a smaller group than he "should," $n_A < n_k$ This reinforces the tendency toward underproduction (A downward and leftward shift in *SS* has the same effect)

2 In his own interest Mr K will not choose a demand-efficient jurisdiction, that is, he will reject point x''_k This reinforces the nonoptimal decision of the old residents

3 Once a suboptimal level of public good is voted, all members have a cost incentive to expand membership to a point such as n''_A

4 As this process continues, a jurisdiction becomes "integrated " Obviously this static example has the makings of a dynamic model of determination of public good supply

Another incentive individuals may have to create a nonoptimal set of jurisdictions, or to maintain the existence of such a nonoptimal set, will arise from any opportunity to restrict the number of jurisdictions If a new entrant to a local jurisdiction pays the marginal cost he imposes on his system, established members of collectives may benefit if they restrict creation of new jurisdictions For by restricting entry they will force new members to join and thereby to overcrowd existing jurisdictions, in this way old residents might turn a profit on their "old" low-cost memberships (capitalized when they sell out) Established members of established local governments may desire to develop and exploit such a monopoly position whenever possible

We can employ figure 7 to show the implications of such monopolistic behavior by *established residents Uniform restrictions on freedom of* creating new jurisdictions would tend to produce "overcrowded" groups as shown by the inefficient supply curve $S''S''$ Clearly, we can repeat for figure 7 the sort of analysis behind figure 6, this analysis again would show that static, nonoptimal collective behavior (restrictions on entry) is reinforced by the individual "dynamic" behavior of the newcomer's choice among the inefficient options Specifically, Mr K in figure 7 will choose a nonoptimal jurisdiction such as A'', rather than K, the (unavailable) optimum Note that static "free-rider" behavior shifts the curve *SS* downward, while static "monopolistic" behavior shifts the curve *SS* to the right One might think that these two tendencies cancel out, but that would be wrong, since in fact they reinforce each other In combination these two types of nonoptimal behavior tend to shift groups from optimal jurisdictions like A to nonoptimal outcomes like A^* (B to B^*, etc.) Thus, the combination of these two sorts of nonoptimal individualism just might result in an "efficient" range of jurisdictionals, sizes, and public-good outputs, but with each and every jurisdiction "inhabited"

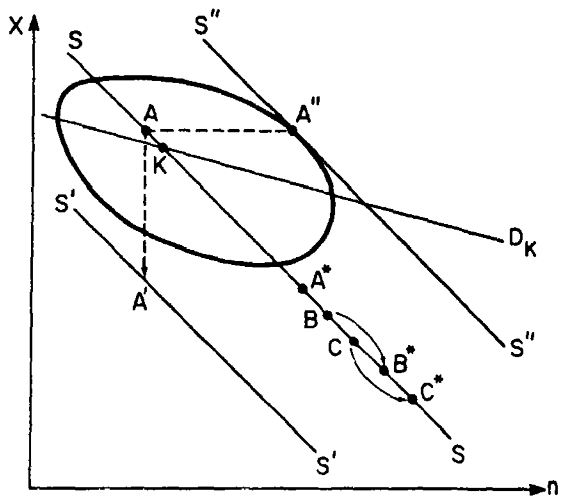


FIG 7

by the "wrong" people *This analysis suggests, therefore, that one key to the maintenance of efficient jurisdictions is a procedure for easily creating new ones*

But is ease of local government formation an unequivocal blessing? I think not Here we may borrow a leaf from Edward Chamberlin (1960), who, in the context of *Monopolistic Competition*, also pondered the costs of diversity in the presence of scale economies [27] Chamberlin showed that when the group of competing monopolists is small—such that each member avoids price-cutting competition—and when at the same time entry to this small group is easy, a stable solution "high up" on each individual producer's average-cost curve may result [28] This notion is obviously applicable to a theory of optimal local collectives, it suggests that a lack of price competition among adjacent groups, when combined with ease of entry, may result *in overdiversification* and overproliferation of small, high-cost jurisdictions

We must close this section of the paper, therefore, with a question The question is, How can one promote price competition among local jurisdictions? The answer is important since it takes us halfway to an efficient configuration of local governments

[27] Roger Noll and Thomas Schelling both suggested this analogy between this model and Chamberlin's Evidently a cost and benefit analysis of diversification has widespread application to economic systems in general, not merely local collectives

[28] See Chamberlin's famous diagram of the individual producer's demand curve sliding up and down an aggregate market demand curve (1960, pp 91–92) The solution I have in mind is the tangency between the aggregate market curve and the single producer's average cost curve

IV. Conclusion

Just what it is that draws some people together and separates others will never be entirely captured by a single theory or model There is no "one thing" which can account for the totality of group formation patterns No claim is made that this paper develops such a universal model But our consolidation of the Tiebout-Buchanan analysis does suggest that one cause may be group formation for the provision of local collective goods and services Even if people are entirely devoid of any feelings toward each other (sympathetic or antipathetic) at a personal level, they may find it in their interests to set up and join associations for the provision of certain collective goods and services—goods which cannot readily be supplied and priced on a variable unit-of-service basis, but which are best provided communally at the same agreed level to all members of the association, in such a way that nonmembers are excluded from enjoying them altogether As a result the same forces which govern the production and consumption of things can lead people into isolated homogeneous groups

In conclusion, we have shown that incentive for segregation may stem from an impersonal, anonymous, unnoticed economic interaction among individuals and technology Such segregation, unlike the forces which call it forth, may be conspicuous, even blatant, and at the same time wholly unintended Ordinarily one might suppose that both kinds of segregation may be present, our kind and Schelling's (see n 1) Where he dealt with a "pure" case in which all the externalities are personal and have to do with people's preferences regarding whom they live among, we have dealt with the opposite "pure" case in which nobody cares in the slightest whom he is with, as long as the commodity value of his "location" yields him the greatest net utility[29] Furthermore, both kinds may often reinforce each other, and undoubtedly a good deal of segregation by color, by language and accent, etc , contains elements of income segregation, of taste segregation, and perhaps then a further reinforcing segregation by sheer class, ethnic, or cultural identity

---

[29] A precautionary remark should be added further to distinguish our model from a world of personal, individualized interaction and conflict First, while the term "local" ordinarily connotes some physical, geographical proximity, a "local public good" in the formal structure of our model may be merely subnational When based on personal tastes for association, segregation is bound to the ideal of physical propinquity and geographical location, whereas the theory of segregated group formation and collective-good provision developed in this paper does not contain an explicit spatial element In principal our model may apply to any good or service which in fact is (or could be) supplied uniformly to a subset of a large (say national) population, the model requires that people make a choice which *identifies* them with a particular version (quality or quantity) of the public good that is available to choose among It is true that such choice most emphatically does arise when people decide to live in the same locale It is also true that the instances of segregation which concern society most often involve geographically concentrated, self-conscious groups Nevertheless, the fact that people are geographically bound plays no explicit role in our analysis, and choice of locale is merely one instance (the most interesting from a practical standpoint) of the larger class of decisions by which people identify themselves with a quantity or quality of public good

References

Buchanan, James "An Economic Theory of Clubs." *Economica* 32 (February 1965) 1–14.

——— "Private Goods as Public Goods." In *The Demand and Supply of Public Goods* Chicago Rand-McNally, 1968

Chamberlin, Edward *Theory of Monopolistic Competition* Cambridge, Mass Harvard Univ Press, 1960

Ellickson, Bryan, "Jurisdictional Fragmentation and Residential Choice." *A E R Papers and Proc* 61 (May 1971) 334–39

McGuire, Martin, "Private Good Clubs and Public Good Clubs Economic Model of Group Formation " *Swedish J Econ* 74 (February 1972) 84–99

McGuire, Martin, and Aaron, Henry "Efficiency and Equity in the Optimal Supply of a Public Good " *Rev Econ and Statis* 60 (February 1969) 31–39

Pauly, Mark. "Optimality, 'Public Goods,' and Local Government A General Theoretical Analysis." *J P E* 78, no 3 (May/June 1970) 572–85 (*a*)

——— "Cores and Clubs " *Public Choice* 6 (Fall 1970) 53–65 (*b*)

Rothenberg, Jerome "Local Decentralization and the Theory of Optimal Government " In *The Analysis of Public Output*, edited by Julius Margolis New York Nat Bur Econ Res , 1970

Samuelson, Paul "Pure Theory of Public Expenditure and Taxation " In *Public Economics* New York St Martin's (for Internat Econ Assoc), 1969

Schelling, Thomas C "Models of Segregation " RM 6014-RC Santa Monica, Calif RAND Corp , 1969

Tiebout, Charles M "A Pure Theory of Local Expenditures " *J P E* 64 (October 1956) 416–24

——— "Economic Theory of Fiscal Decentralization " In *Public Finances, Needs, Sources and Utilization* Princeton, N J Nat Bur Econ Res , 1961

Williams, Alan "The Optimal Provision of Public Goods in a System of Local Government " *J P E* 74, no 1 (February 1966) 18–33.

Stability and the Quantity Theory

P W Howitt

Northwestern University and University of Western Ontario

This paper presents a method for analyzing the microdynamics of monetary exchange The central feature of the analysis is that transactions take place during the process of price adjustment The method is used to demonstrate the stability of a long-run equilibrium in which the quantity theory is valid If all goods are gross substitutes and the real balance effect is operative, then the equilibrium is stable if (i) there is no rationing, (ii) there is a form of competitive rationing, or (iii) there are no distribution effects.

The neoclassical long-run quantity theory consists of two basic propositions The first is the familiar quantity theorem—that in equilibrium the level of money prices will be directly proportional to the quantity of money. The second proposition, without which the quantity theorem would have little empirical significance, is that this equilibrium is stable

Patinkin (1956) showed that the neoclassicals had no satisfactory explanation of the stability proposition He suggested that such an explanation could be provided by considering the real balance effect on excess demands If there were no distribution effects, and if appropriate restrictions were placed on the substitutability of commodities,[1] it could then be shown that a Walrasian tâtonnement would converge on a vector of money prices proportional to the stock of money Archibald and Lipsey (1958) provided a method for extending this analysis to the case of

---

This paper is a revised version of one chapter of my doctoral dissertation, which was written with financial support from the Canada Council I am indebted to Robert Clower, John Ledyard, Axel Leijonhufvud, and Joseph Ostroy for helpful comments and discussions

[1] See Patinkin 1956, pp 342–44 Arrow and Hurwicz (1958) showed that a sufficient condition is for all goods to be gross substitutes

distribution effects.[2] If we disturb a long-run equilibrium by, say, doubling the quantity of money, the temporary equilibrium established by a tâtonnement might contradict the quantity theorem unless each transactor's money balances are exactly doubled. However, these distribution effects are transitory Money balances will change hands from one equilibrium to the next, until in the limit each transactor holds exactly twice his original money holdings and the money price of each commodity exactly doubles

As amended by Archibald and Lipsey, Patinkin's analysis does provide an explanation of the stability proposition It is in fact the only such explanation in the literature However, the analysis has the well-known drawback of Walras's original formulation of the tâtonnement process—that no transactions occur until market clearing prices have been established To state the issue in familiar language, our understanding of the quantity theory is still impaired by a dichotomy There are two basic forces that act to propel the economy from one long-run equilibrium to the next First, there is the force of competitive bidding that makes prices adjust to their market clearing values Second, there is the force of ongoing transactions that makes each transactor's money holding tend to its stationary level While it is commonly acknowledged that the quantity theory is meant to apply to a world where these forces act simultaneously, the most complete explanation available of the quantity theory requires that we consider them separately—while the one is operating the other must be suspended [3]

Clower (1965) and Leijonhufvud (1968) have demonstrated that the basis of Keynes's attack on orthodox monetary theory was the contention that this dichotomy was invalid They show how false trading gives rise to the multiplier process that can make the economy move away from its full-employment equilibrium, even with flexible wages and prices If the quantity theory is to be maintained in the face of the Keynesian challenge, or if we are to have a monetary theory that can in fact reconcile Keynesian economics *with the quantity theory, we need some method for presenting* the stability proposition that does not rely on the dichotomy of price adjustment and transactions The purpose of this paper is to present such a method

In order to do away with this dichotomy we must reexamine the microfoundations of the quantity theory This is because, first, there is nothing in accepted general equilibrium theory to tell us how transactions occur even in equilibrium,[4] much less when there are nonzero excess demands. Because the time path of prices is affected by the false trading[5] it becomes important to determine what transactions take place. Second, a meaningful monetary theory must specify a money commodity that changes hands with each transaction, this is not done by accepted general equilibrium theory [6] Third, the excess-demand functions of general equilibrium theory are usually derived under the implicit assumption that all planned trades will in fact be made, an assumption which appears to be inconsistent with the existence of false trading [7]

---

[2] Archibald and Lipsey's analysis was the subject of the symposium by Baumol et al (1960) Their model was extended by the exchange between Hahn (1962) and Clower (1963) to include a richer description of intertemporal choice

[3] Grossman's (1971) interesting analysis of price dynamics in a Patinkinesque model with false trading fails to do away with the dichotomy because prices do not change until quantities have all adjusted to a "fix-price" equilibrium, the effect is to retain the dichotomy by reversing the relative speeds of price adjustment and transactions The same is true of Frevert (1970)

To deal with these issues we shall describe a highly simplified exchange system organized by a group of shopkeepers One of the consequences of this approach is that we are able to handle these problems, which arise in Patinkin's Walrasian approach, with a model that is formally very similar to Patinkin's Indeed the long-run equilibria of the two models are identical This allows us to demonstrate the stability proposition using many of the tools developed for the Walrasian system

I. The Exchange Environment

In order to keep track of actual transactions, our description of the shopkeeper system will be very explicit about how these transactions occur There are in the economy B transactors and $C + 1$ commodities For reasons that will soon be apparent, the $(C + 1)$st commodity is called "money " Each transactor b begins each period with the endowment $W_c^b \geq 0$ of commodity c, $c = 1, \quad, C$, and an endowment $M_b \geq 0$ of money The first C commodities are nondurable goods that disappear after each period [8] These endowments reappear at the beginning of each period in unchanging quantities There are C shops in the economy In shop c the only exchanges that may be made are trades of money and commodity c for each other Each shop is presided over by a separate shopkeeper whose job it is to set the price for his commodity and to facilitate the exchange of that commodity from seller to buyer by acting as middleman

Trading proceeds once a period in each shop During the "morning," shopkeeper c posts a money price, p_c, that will prevail for all exchanges on that day Later in the day each transactor goes to all the shops where he wishes to do business, placing an order to buy or sell a specific quantity of each commodity. We denote by X_c^b the bth transactor's order for commodity c

---

[4] This point is made very forcibly by Ostroy (1973).

[5] See Edgeworth (1881, pp 22–30) and Kaldor (1934)

[6] See Clower 1967

[7] See Clower 1965, p 117.

[8] Even if money were not the only durable commodity it would still have the distinguishing feature of being the unique medium of exchange (see below). However the nondurable nature of nonmoney commodities means that they will not be demanded for speculative purposes.

In order to make money a well-defined medium of exchange we suppose that the shopkeepers require all purchase orders to be accompanied by the total money payment This requires that·[9]

$$\sum_{c \in P(b)} p_c X_c^b \leq M_b \text{ for all } b, \tag{1}$$

where

$$P(b) = \{c \quad X_c^b \geq 0\}$$

For a sell order the transactor must deposit with the shopkeeper the entire quantity offered for sale This requires that

$$X_c^b \geq -W_c^b \text{ for all } b \text{ and } c \tag{2}$$

These conditions ensure that the transactor can only offer to trade commodities that are available at the beginning of the period This rules out forward transactions and eliminates the possibility of bankruptcy [10]

The sum of all orders for commodity c,

$$X_c = \sum_{b=1}^{B} X_c^b,$$

is called the "excess demand" for commodity c

Once all orders are placed the shopkeepers have two tasks First they convert orders into transactions If commodity c has a zero excess demand, we assume all orders for commodity c are executed [11] If excess demand is not zero, the shopkeeper has some rule to determine what transactions occur The quantity of commodity c the shopkeeper for that commodity allows transactor b to exchange is T_c^b, which is positive (negative) for a purchase (sale) The shopkeeper's transaction rules make T_c^b a function of all orders for commodity c and the vector of money holdings [12] That is $T_c^b = T_c^b(X_c^1, \quad , X_c^B, M)$ for all b and c, where $M = (M_1, \quad , M_B)$ We assume that for all b and c, T_c^b is a continuous function of all arguments, satisfying a Lipschitz condition on every compact set, and

$$T_c^b = X_c^b \qquad \text{if } X_c = 0,$$

$$|T_c^b| \leq |X_c^b|$$

and

$$T_c^b \cdot X_c^b \geq 0$$

---

[9] This condition is similar to Clower's (1967) "dichotomized budget constraint "

[10] The possible complications of bankruptcy have been stressed by Hahn (1965)

[11] This rules out the sort of phenomenon that Phelps (1970) described with job vacancies and involuntarily unemployed workers at the same time

[12] The reason for including money balances will become apparent in Section IV, in short, it allows us to include the competitive-rationing restrictions (see below)

The first of these conditions has already been explained. The others say that no orders will be overexecuted and no purchase (sale) order will be converted into a sale (purchase) It follows from the last two that if no order is placed, no transaction will occur

In the "afternoon" the shopkeepers allow these transactions to occur and return the deposits corresponding to unexecuted orders Thus the transactors' money balances for the beginning of the next period are given by the difference equations

$$\Delta M_b = -\sum_{c=1}^{C} p_c T_c^b \qquad \text{for all } b\,^{13} \tag{3}$$

The shopkeeper's second task is to determine next period's price. We assume he follows the usual adjustment rule of raising (lowering) his price when there is a positive (negative) excess demand That is,

$$\Delta p_c = H_c \qquad \text{for all } c, \tag{4}$$

where

$$H_c \gtreqless 0 \quad \text{as} \quad X_c \gtreqless 0.$$

These tasks of the shopkeepers are not meant to be literal accounts of all exchange and bargaining activities in a monetary economy Instead they are intended to personify the "forces of competition " In this respect the shopkeepers are much like the Walrasian auctioneer. When X_c is positive there will be some buyers who, in the absence of the shopkeeper, would be unable to find a supplier of commodity c In a Walrasian tâtonnement all trading would be suspended, some frustrated buyers would suggest that trading resume with a higher p_c, and orders would again be submitted on the basis of the new price The auctioneer is supposed to embody these forces acting on prices In the environment outlined above, the bargaining cannot be suspended in the middle of a period Excess demands become apparent to the transactors only after trade has taken place, only then can frustrated buyers and sellers make known their willingness to trade at different prices. We assume that the shopkeepers have been provided with some incentive to reflect these competitive forces in their pricing decisions

II. Excess-Demand Functions

Within this shopkeeper system we can derive excess-demand functions similar to Patinkin's. However, the possibility of false trading means these functions cannot be derived in the usual way because the typical transactor's transactions depend not only upon his orders but also upon how the shopkeepers convert orders into transactions. Although shopkeepers

---

[13] The Δ represents a first difference Time subscripts are omitted here.

have deterministic transaction rules, the relationship between a transactor's orders and transactions will appear to him as being stochastic because he does not have all the information available to the shopkeepers.

One implication of this false trading is that the transactor's current sales income can have a direct effect on his demand orders, which cannot occur in the general equilibrium theory of Patinkin with no false trading [14] With no false trading, a change in one transactor's demands can only affect another transactor's demands if it somehow results in a price change With false trading, a reduction in some transactor's demand for one commodity can cause sellers of that commodity to reduce their demands for other commodities, even at unchanged prices, because their current sales income will have fallen This intermarket connection that arises from false trading has been interpreted by Clower (1965) and others[15] as the theoretical essence of the Keynesian consumption function

Another implication of false trading is that the typical transactor will be especially concerned with the state of his money balances Money provides him with a form of insurance against the possibility that sell orders might not all be executed in the future, which would force a reduction in purchase orders if sufficient money balances were not held Even if some commodities other than money were durable, money would still be especially important to the transactor because of the constraint (1) that assigns the role of medium of exchange to only the one commodity, money Because money is the only medium of exchange it is also the only commodity that can provide that particular form of insurance Even if current sales income were to have no direct effect on excess demands this special role of money would provide the same sort of intermarket connection as explained above An exogenous decrease in demand for one commodity will reduce the money holdings of sellers of that commodity, providing them with an incentive to reduce their demands for other commodities even with no change in prices [16]

To be perfectly explicit about all this we would characterize the transactor's decision problem as the maximization of the expected value of an intertemporal utility function, where actual transactions depended upon orders placed but with random error terms to account for the shopkeepers' activities Money balances would not enter this utility function directly but would be involved in the constraint (1), which would have to hold in every time period Money would become the "state variable" of the decision problem To include the direct effect of current sales income we could assume that the transactor revised his estimates of the probability distribution of the error term every time a new observation was generated by the shopkeepers If a reduction in demand for a commodity he was selling caused a decrease in his sales without an initial change in his sell order, he would form more pessimistic expectations about how the shopkeeper for that commodity was rationing him This would cause him to reduce his demands elsewhere because of a reduction in his future sales prospects[17]

---

[14] This point is made very forcibly by Clower (1965, pp 111–12)

[15] Barro and Grossman 1971, and Leijonhufvud 1968

[16] I have shown elsewhere (1973) how this intermarket connection can produce, in a model like the present one, the sort of Keynesian short-run behavior usually associated with current income having an important effect on excess demands in the form of the consumption function

Rather than describe the intertemporal decision problem explicitly we proceed in the following way We make the simplifying assumption that the transactor has no idea how the future relationship between orders and transactions will be affected by variables about which he has any information, including the current decisions of the shopkeepers In this case the subjective probability distribution of error terms is an unchanging one, and there is no longer any direct effect of current sales income To incorporate the direct effect of money balances, we characterize the decision process as a single period choice, where the choice variables are this period's orders and the money balances the transactor would have at the end of the period if all orders were executed The bth transactor's objective function is $U^b(X^b, M_b + m_b, p)$, where $X^b = (X_1^b, \quad , X_C^b)$, $p = (p_1, \quad , p_C)$, and the intended change of M_b over the period is

$$m_b = -\sum_{c=1}^{C} p_c X_c^b \tag{5}$$

Prices enter into this objective function because the usefulness of money depends upon its future purchasing power We make the usual homogeneity assumption with respect to all nominal variables, namely, that[18] $U^b(X^b, M_b + m_b, p) = U^b(X^b, \lambda M_b + \lambda m_b, \lambda p)$ for all $\lambda > 0$ This is equivalent to saying that real instead of just nominal balances are considered important The most convenient rationalization is to assume that all future price expectations are unit elastic with respect to current prices[19] Therefore if p, m_b, and M_b were all to double, all future price expectations would double, and the future purchasing power of the transactor's end of period money balances would be unaffected

---

[17] This way of accounting for the direct effect of sales income differs from Clower's 1965) "dual decision hypothesis" used by Barro and Grossman (1971) and Grossman 1971) The difference is that for these writers sales income during the current period is nown before demand orders are placed This is equivalent to making expected income lways equal to current income, implying a much closer relationship between income and onsumption than implied by modern consumption theories This point has been made y Leijonhufvud (1973) The present analysis would allow for a "looser" dependence of onsumption on income

[18] See Samuelson 1947, p 119

[19] See Hicks 1946, p. 205.

The transactor's decision problem is to choose X^b and m_b so as to maximize U^b subject to (1), (2), and (5) The solution to this decision problem implicitly defines the transactor's net demand functions.

$$X^b = X^b(p, M_b) = (X_1^b(p, M_b), \ldots, X_C^b(p, M_b))$$

$$m_b = m_b(p, M_b) = -\sum_{c=1}^{C} p_c X_c^b(p, M_b).$$

We assume X^b and m_b possess continuous first-order partial derivatives. The net demands may be summed to form the excess-demand function

$$X_c(p, M) = \sum_{b=1}^{B} X_c^b(p, M_b)$$

This is the same formal model of the transactor's decision process as Archibald and Lipsey's (1958) except for the constraint (1). This constraint makes the model very similar to Clower's (1967) [20] Including the constraint allows us to avoid the anomalous results found by Clower (1967) and Hahn (1965) in Patinkin's general equilibrium theory—that the economic system could proceed as usual even without the medium of exchange because there is no peculiar characteristic of money that would require any transactor to hold positive quantities of it in equilibrium This anomalous result is avoided with the constraint (1) because there is now a compelling reason to hold money—a transactor needs to have some at the end of this period in order to submit purchase orders at the beginning of the next period Although we have derived this demand for money partly from putting money into the utility function this is not subject to the usual criticism that it begs the question of why transactors hold money they hold it in order to place purchase orders

III. The Dynamical System

We may complete the description of price adjustment given in (4) by assuming that there exist continuously differentiable functions $H_c(p, M)$, for all c, such that $H_c = H_c(p, M)$ If we approximate the time paths generated by (3) and (4) by their continuous time analogs, we get the following system [21]

$$\dot{p}_c = H_c(p, M), c = 1, \ldots, C,$$

$$M_b = -\sum_{c=1}^{C} p_c T_c^b[X_c^1(p, M_1), \ldots, X_c^B(p, M_B); M]; b = 1, \ldots, B.$$

---

[20] Clower's excess demand functions have been analyzed by Lloyd (1971) and Clower (1971)

[21] The "dots" refer to time derivatives

We refer to this system as Q. An equilibrium to Q is a vector $(p, M) > 0$[22] such that the right-hand side of Q equals zero. We assume that at least one such equilibrium exists In equilibrium excess demands are all zero and each transactor's money holding is stationary This is the long-run equilibrium analyzed by Archibald and Lipsey (1958).

The two stages of the Patinkin, Archibald-Lipsey process are merged into one by Q. The tâtonnement stage could be represented by holding M constant and allowing prices to adjust If this were stable it would establish a temporary equilibrium p with $X_c(p, M) = 0$ for all c The second stage could be represented by assuming the tâtonnement to take place instantaneously, so we are always in a temporary equilibrium. The adjustment of money balances would then be given by $\dot{M}_b = m_b(p, M_b)$ because there would be no rationing in the temporary equilibrium

The system Q satisfies the following two properties

1 Walras' law

$$\sum_{c=1}^{C} p_c X_c(p, M) + \sum_{b=1}^{B} m_b(p, M_b) = 0;$$

2 Homogeneity of net demands For any $\lambda > 0$,

$$X_c^b(p, M_b) = X_c^b(\lambda p, \lambda M_b), \quad \text{and} \quad m_b(p, M_b) = \frac{1}{\lambda} m_b(\lambda p, \lambda M_b)$$

In proving the stability proposition we make the following two assumptions

1 Gross substitutability

$$\frac{\partial}{\partial p_k} X_c^b(p, M_b) > 0 \quad \text{and} \quad \frac{\partial}{\partial p_k} m_b(p, M_b) > 0, \qquad \text{for all } k, b, c(c \neq k),$$

2 Real balance effect

$$\frac{\partial}{\partial M_b} X_c^b(p, M_b) > 0 \qquad \text{for all } b, c$$

The system Q together with gross substitutability and real balance effect is referred to as Q^* There is a very striking resemblance between Q and the Walrasian system Some of the results which have been developed in the Walrasian framework carry over directly to Q, and we make use of this strong analogy in proving the quantity theorem and the stability proposition.

---

[22] We use the following notation throughout for any n-dimensional vectors, $x = (x_1, \ldots, x_n)$ and $x' = (x_1', \ldots, x_n')$; $x > x'$ means $x_i > x_i'$ for all i, $x \geq x'$ means $x_i \geq x_i'$ for all i, and $x \geq x'$ means $x \geq x'$ and $x \neq x'$.

Before proving the quantity theorem we state a result very similar to a lemma proved by Arrow, Block, and Hurwicz (1959, p. 88).

ABH lemma: Given gross substitutability and the real balance effect, the relations $0 < (p', M') \leq (p'', M'')$, $p'_c = p''_c$ for $c \in R \subset \{1, \ldots, C\}$, $M'_b = M''_b$ for $b \in S \subset \{1, \ldots, B\}$; imply that (*a*) $X_c(p', M') < X_c(p'', M'')$ for all $c \in R$, and (*b*) $m_b(p', M'_b) < m_b(p'', M''_b)$ for all $b \in S$ if and only if $p' \neq p''$.

We may now prove a strong version of the quantity theorem.[23]

Quantity theorem: Given any value $\mathbf{M} > 0$ of the money supply, there is exactly one equilibrium to Q^* for which

$$\sum_{b=1}^{B} M_b = \mathbf{M}$$

Given two values of the money supply, $\mathbf{M}$ and $\mathbf{M}'$, if (p, M) and (p', M') are the corresponding equilibria, then $(p', M') = (\beta p, \beta M)$ where $\beta = (\mathbf{M}'/\mathbf{M})$.

Proof: To prove uniqueness for a given $\mathbf{M}$, suppose (p, M) is an equilibrium to Q^* such that

$$\sum_{b=1}^{B} M_b = \mathbf{M}$$

Take any other vector, $(\bar{p}, \bar{M})$ such that $(\bar{p}, \bar{M}) \neq (p, M)$, and such that

$$\sum_{b=1}^{B} \bar{M}_b = \mathbf{M}$$

We show that $(\bar{p}, \bar{M})$ cannot be an equilibrium. Define $\lambda = \min\{\bar{p}_1/p_1, \ldots, \bar{p}_C/p_C, \bar{M}_1/M_1, \ldots, \bar{M}_B/M_B\}$. Since

$$\sum_{b=1}^{B} \bar{M}_b = \sum_{b=1}^{B} M_b,$$

it cannot be that $(\bar{p}, \bar{M}) = (\lambda p, \lambda M)$, for this would imply that $(p, M) = (\bar{p}, \bar{M})$. Therefore, $(\lambda p, \lambda M) \leq (\bar{p}, \bar{M})$. There are two cases to consider.

a) If there is some c such that $\lambda = \bar{p}_c/p_c$, then by part (*a*) of the ABH lemma and by homogeneity, $0 = X_c(p, M) = X_c(\lambda p, \lambda M) < X_c(\bar{p}, \bar{M})$. Therefore $(\bar{p}, \bar{M})$ is not an equilibrium.

b) If there is no such c, then $\lambda p \neq \bar{p}$, and there is some b such that $\lambda = \bar{M}_b/M_b$. Therefore, by part (*b*) of the ABH lemma and by homogeneity, $0 = \lambda m_b(p, M_b) = m_b(\lambda p, \lambda M_b) < m_b(\bar{p}, \bar{M}_b)$. Therefore, $(\bar{p}, \bar{M})$ is not an equilibrium.

---

[23] This proof is similar to the proof of the analogous uniqueness theorem of Arrow et al. (1959, pp. 89–90).

The rest of the proof follows directly from the homogeneity of net demands. If (p, M) is an equilibrium to Q^* such that

$$\sum_{b=1}^{B} M_b = \mathbf{M},$$

and if $\beta = (\mathbf{M}'/\mathbf{M})$, then by homogeneity $X_c(\beta p, \beta M) = X_c(p, M) = 0$, for all c, and $m_b(\beta p, \beta M_b) = \beta m_b(p, M_b) = 0$ for all b Therefore $(\beta p, \beta M)$ is an equilibrium to Q^* and

$$\sum_{b=1}^{B} \beta M_b = \mathbf{M}'$$

It follows from the previous paragraph that $(\beta p, \beta M)$ is the only such equilibrium for M'

This is the quantity theorem in its fullest sense If, starting from a long-run equilibrium, the stock of money is doubled in the economy, then, regardless of how this money is initially distributed, the economy will not come to rest again until all prices have exactly doubled and every transactor's money holdings have exactly doubled [24] All that is missing from the quantity theory as stated in the introduction to this study is a proof that the economy would indeed come to rest if it were disturbed in this or in any other way

IV. The Stability Proposition

We interpret the stability proposition to mean that Q is asymptotically stable That is, for any initial position $(p^o, M^o) > 0$, every solution $(p(t, p^o, M^o), M(t, p^o, M^o))$ to Q starting at (p^o, M^o) converges to an equilibrium as $t \rightarrow \infty$ We assume that for any initial position a unique solution exists to Q, which is continuous with respect to the initial position and which is bounded [25] We can demonstrate that Q^* is asymptotically stable under three different sets of restrictions, as in the following three theorems

For the first two theorems we shall use the result, due to Uzawa (1961, p 619), that Q^* is quasi-stable[26] if there exists a continuous function,

---

[24] We have been unable to find in the literature a proof that the long-run quantity heory equilibrium of stocks and prices is unique Patinkin (1965, p. 45) *assumed* the iniqueness of equilibrium in order to derive the quantity theorem. The same is true of Archibald and Lipsey (1958). Arrow et al. (1959, pp. 89–90) proved the uniqueness of prices for given stock holdings Hadar (1965) proved the uniqueness of stock holdings or given prices.

[25] I have elsewhere (1973) shown that the system Q^* satisfies all three assumptions under the premises of theorems 1–3 below.

[26] The system Q^* is defined to be quasi-stable if for any initial position $(p^o, M^o) > 0$ any solution to Q^* is bounded and every limit point of the solution is an equilibrium.

$V(p, M)$, called a "modified Lyapunov function," defined on R_{+}^{B+C},[27] such that for every $(p^o, M^o) > 0$ the function $v(t) = V[p(t; p^o, M^o), M(t, p^o, M^o)]$ is a strictly decreasing function with respect to t unless $p(t; p^o, M^o), M(t, p^o, M^o)$ is an equilibrium to Q^*.

First, suppose the shopkeepers can maintain stocks of the commodities and do not ration the transactors. All excess demands are absorbed by the shopkeepers' inventories. In this case we can use the modified Lyapunov function. $V_1(p, M) = \max J(p, M) - \min J(p, M)$, where $J(p, M) = \{p_1/\bar{p}_1, \ldots, p_C/\bar{p}_C, M_1/\bar{M}_1, \ldots, M_B/\bar{M}_B\}$, and $(\bar{p}, \bar{M})$ is an equilibrium to Q. An intuitive sketch of this stability proof is helpful. If $\min J(p, M) = p_c/\bar{p}_c$, then p_c is so low that $X_c > 0$ and therefore $\dot{p}_c > 0$. If $M_b/\bar{M}_b = \min J(P, M)$, then M_b is so low that $m_b > 0$. Because there is no rationing, the bth transactor's attempt to increase his money holding will be successful, and $\dot{M}_b > 0$. Similarly, the maximum of the ratios will be decreasing.

Theorem 1. If $T_c^b \equiv X_c^b$, Q^* is asymptotically stable.

Proof. In this case $\dot{M}_b(t) = m_b[p(t), M_b(t)]$ for all b.[28] We define

$$\bar{v}_1(t) = \max J[p(t), M(t)],$$

$$\mathbf{v}_1(t) = \min J[p(t), M(t)],$$

$$v_1(t) = \bar{v}_1(t) - \mathbf{v}_1(t),$$

$$w_1(t) = \limsup_{d \to 0} [v_1(t + d) - v_1(t)]/d,$$

$$\bar{w}_1(t) = \limsup_{d \to 0} [\bar{v}_1(t + d) - \bar{v}_1(t)]/d,$$

$$\mathbf{w}_1(t) = \liminf_{d \to 0} [\mathbf{v}_1(t + d) - \mathbf{v}_1(t)]/d,$$

$V_1(p, M)$ is continuous on R_{+}^{B+C}. Take any t such that $(p(t), M(t))$ is not an equilibrium. To prove $v_1(t)$ is strictly decreasing we show that $w_1(t) < 0$. It can be shown that

$$1)\quad w_1(t) \leq \bar{w}_1(t) - \mathbf{w}_1(t).$$

First consider $\mathbf{w}_1(t)$. It can be shown[29] that (a) there is some c such that $\mathbf{v}_1(t) = p_c(t)/\bar{p}_c$ and $\mathbf{w}_1(t) = \dot{p}_c(t)/\bar{p}_c$, or (b) there is some b such that $\mathbf{v}_1(t) = M_b(t)/\bar{M}_b$ and $\mathbf{w}_1 = \dot{M}_b(t)/\bar{M}_b$. If (a) then, defining $\lambda = \mathbf{v}_1(t)$, $(p(t), M(t)) \geq (\lambda\bar{p}, \lambda\bar{M})$ and $p_c(t) = \lambda\bar{p}_c$. Therefore, by homogeneity and the ABH lemma, $X_c[p(t), M(t)] > X_c(\lambda\bar{p}, \lambda\bar{M}) = 0$, therefore $\mathbf{w}_1(t) > 0$. If (b) and $p(t) \neq \lambda\bar{p}$, then $m_b[p(t), M_b(t)] > m_b(\lambda\bar{p}, \lambda\bar{M}_b) = 0$; therefore

---

[27] For any integer n, R_{+}^{n} is the set of all n-dimensional vectors x such that $x > 0$.

[28] The arguments (p^o, M^o) are omitted whenever no ambiguity results.

[29] Arrow et al. 1959, pp. 96–97.

$\mathbf{w}_1(t) > 0$. If (b) and $p(t) = \lambda\bar{p}$, then $m_b[p(t), M_b(t)] = m_b(\lambda\bar{p}, \lambda\bar{M}) = 0$, and $\mathbf{w}_1(t) = 0$ It follows that

ii) $\mathbf{w}_1(t) \geq 0$ with equality only if $\mathbf{v}_1(t)\bar{p} = p(t)$

By an analogous argument,

iii) $\bar{w}_1(t) \leq 0$ with equality only if $\bar{v}_1(t)\bar{p} = p(t)$

The equality cannot hold in both (ii) and (iii) because this would imply $\mathbf{v}_1(t) = \bar{v}_1(t)$, in which case $(p(t), M(t))$ would be an equilibrium Therefore, combining (i), (ii), and (iii), $w_1(t) < 0$, and hence Q^* is quasi-stable It can also be shown[30] that $\bar{v}_1(t) - \mathbf{v}_1(t) \rightarrow 0$ as $t \rightarrow \infty$, so that $(p(t), M(t)) \rightarrow (\alpha\bar{p}, \alpha\bar{M})$ for some $\alpha > 0$ Therefore Q^* is asymptotically stable

Next, suppose that at least some rationing does take place out of equilibrium We can demonstrate the stability proposition again, provided the transaction rules satisfy a restriction that we shall now describe It is clear from the proof of theorem 1 that Q^* will also be stable if for all b, $T_c^b = X_c^b$ for $X_c^b \geq 0$ whenever $M_b(t)/\bar{M}_b \geq M_k(t)/\bar{M}_k$ for all k, and $T_c^b = X_c^b$ for all $X_c^b \leq 0$ whenever $M_b(t)/\bar{M}_b \leq M_k(t)/\bar{M}_k$ for all k This restriction is referred to as "competitive rationing" Given competitive rationing then, referring to our intuitive sketch of theorem 1, if $M_b/\bar{M}_b = \min J(p, M)$, then transactor b will not be rationed on sell orders and $M_b \geq m_b > 0$ Similarly if $M_b/\bar{M}_b = \max J(p, M)$, then $M_b \leq m_b < 0$.

We call this restriction competitive for two reasons. First, it says that the shopkeepers must assign the most favorable queueing position to those who have the greatest incentive to be first in line, in the sense that a transactor with low money balances will want to sell more than the same transactor with high money balances Second, we generally think of the forces of competition underlying the price adjustments as being the bids and offers made by those who have been unsuccessful in trading at the previous prices Their competitive bidding should have the effect not only of changing prices but also of allowing those who have been rationed most in the past to be rationed least in the present, because unless there are barriers to entry, the unsuccessful seller has only to underbid his competitutors in order to sell his goods[31] If the transactor with the "lowest" money balances is also the one whose sell orders have been most heavily rationed, the competitive rationing restriction can be interpreted as reflecting his successful competitive bidding Thus we have the second version of the stability proposition.

---

[30] Uzawa 1961, p. 624

[31] This phenomena would be more easily formalized in a model where each transactor is allowed to set his own bid or asked price, which in turn would determine his share of the rationing This is the approach of Bushaw and Clower (1957, pp 176–90) and Fisher (1972)

Theorem 2 Given competitive rationing, Q^* is asymptotically stable

For the final version, suppose the shopkeepers hold no inventories at all, in which case

$$\sum_{b=1}^{B} T_c^b = 0$$

for all c With no inventories we cannot be sure that there exists a set of transaction rules satisfying competitive rationing Even the transactor who is first in line might have to be rationed if, say, his net demand is negative and exceeds in absolute value the sum of all positive demands for that commodity Therefore the stability proposition cannot be demonstrated using the same method as in theorems 1 and 2 However, it can be demonstrated for the case of no distribution effects, that is, $X_c(p, M) = X_c(p, M')$ for all p, M, and M' such that

$$\sum_{b=1}^{B} M_b = \sum_{b=1}^{B} M_b'$$

Theorem 3 Given no distribution effects, if

$$\sum_{b=1}^{B} T_c^b = 0$$

for all c then Q^* is asymptotically stable

Proof Take any $(p^o, M^o) > 0$ Define

$$\mathbf{M} \equiv \sum_{b=1}^{B} M_b^o$$

Let $(\bar{p}, \bar{M})$ be the unique equilibrium such that

$$\sum_{b=1}^{B} \bar{M}_b = \mathbf{M}$$

Because

$$\sum_{b=1}^{B} T_c^b = 0$$

for all c, we have

$$\frac{d}{dt}\sum_{b=1}^{B} M_b = -\sum_{b=1}^{B}\sum_{c=1}^{C} p_c T_c^b = 0$$

Therefore for all $t \geq 0$

$$\sum_{b=1}^{B} M_b(t) = \mathbf{M}$$

and, because of no distribution effects, $X_c[p(t), M(t)] = X_c[p(t), \bar{M}]$ for all c Suppose $(p(t), \bar{M})$ is not an equilibrium Then by the ABH lemma, if $p_c(t)/\bar{p}_c = \min \{p_k(t)/\bar{p}_k\}$ then $X_c[p(t), M(t)] = X_c[p(t), \bar{M}] > X_c(\bar{p}, \bar{M}) = 0$, and $\dot{p}_c(t) > 0$ Similarly if $p_c(t)/\bar{p}_c = \max \{p_k(t)/\bar{p}_k\}$ then $\dot{p}_c(t) < 0$ It follows[32] that (i) $p(t) \to \bar{p}$ as $t \to \infty$. Because of (i) and the continuity of X_c it follows that for all c, $X_c[p(t), M(t)] = X_c[p(t), \bar{M}] \to X_c(\bar{p}, \bar{M}) = 0$ as $t \to \infty$ Therefore, because T_c^b and X_c^b are continuous, and because $T_c^b = X_c^b$ whenever $X_c = 0$, for all b $M_b(t) \to m_b[p(t), M_b(t)]$ as $t \to \infty$. It follows from (i) and the continuity of m_b that for all $b m_b[p(t), M_b(t)] \to m_b[\bar{p}, M_b(t)]$ as $t \to \infty$ Therefore the motion of M_b for any b is given by the nonautonomous system $M_b(t) = m_b[\bar{p}, M_b(t)] + h_b(t)$, where $h_b(t) \to 0$ as $t \to \infty$ The corresponding autonomous system, $M_b = m_b[\bar{p}, M_b(t)]$, is asymptotically stable because, from the real balance effect, $(d/dt)(M_b - \bar{M}_b)^2 = 2(M_b - \bar{M}_b)m_b(\bar{p}, M_b) < 0$ Therefore[33] the nonautonomous system is also stable and, (ii) $M_b(t) \to \bar{M}_b$ for all b as $t \to \infty$ The conclusion follows from (i) and (ii)

V. Relationship to the Nontâtonnement Literature

Mathematical economists who have studied the dynamics of nonrecontracting exchange models have not had to impose gross substitutability on the excess-demand functions in order to prove the convergence of prices and commodity holdings [34] This has given rise to the conjecture that the recontracting itself was a source of instability—that by permitting false trades an unstable exchange model could be made stable [35] The question naturally arises why we had to impose gross substitutability in the present model

There are two fundamental differences between the present model and those of the "nontâtonnement" literature Both of these rule out the stability proofs that have been used in that literature First, prices enter the utility functions in the present model but not in the nontâtonnement literature Second, the nontâtonnement literature has been concerned exclusively with pure stock models, whereas we are dealing here with a stock-flow model

Uzawa's (1962) method of proof was to use the negative of the sum of all transactors' utilities as a modified Lyapunov function This function will be decreasing over time in his model if every transactor refuses to make any trade that results in a decrease in utility. However, once prices enter the utility function a transactor's refusal to accept a utility-decreasing trade will no longer guarantee that his utility will be increasing over time, for the price changes can make it decrease. There appears to be no way of avoiding this in a model of monetary exchange where money is only held with a view to its future purchasing power Also, in a stock-flow model where the flow commodities disappear after the trading period the transactor's utility might be decreasing if he is unable to engage in the same trades as in the previous period, even if he refuses to engage in utility-decreasing trades

---

[32] Uzawa 1961, pp 624

[33] Coddington and Levinson 1955, p 327

[34] The two most important works in this area are those of Hahn and Negishi (1962) and Uzawa (1962) The literature is summarized by Negishi (1962)

[35] This conjecture is referred to (although not necessarily maintained) by Negishi (1962, pp. 658–59) and Arrow and Hahn (1971, p 328)

Hahn and Negishi's (1962) proof depends upon an assumption that the value of this period's endowments for any transactor, evaluated at last period's prices, is the same as last period's endowment evaluated at last period's prices In a pure stock model this assumption is natural because this period's endowments are all the result of last period's trades, none of which can change the value of the transactor's commodity holdings However, this assumption cannot be carried over to a stock-flow model where production and consumption also change the endowments Arrow and Hahn (1971) and Fisher (1972) have transposed Hahn and Negishi's analysis to a model of monetary exchange, but they too have a pure stock analysis without prices in the utility function

VI. Summary and Conclusion

We have presented a method by which the stability proposition of the quantity theory can be verified without the artificial separation of exchange and price adjustment The basic features of the formal model were that sales income was not an argument of the excess-demand functions, that all commodities were gross substitutes, and that the real balance effect was operative on all transactors' net demands It was shown that the long-run equilibrium of the quantity theory was asymptotically stable if there was no rationing, if there was a form of "competitive rationing," or if there were no distribution effects

Of course it does not follow from acceptance of this stability proposition that one must then accept the policy recommendations of the quantity theorists, even if the proposition were to be proved in a much more realistic model than the one used here The main reason for this is that the stability proposition refers only to the behavior of an unchanging system as time goes to infinity. We still have no well-articulated theory to tell us what to expect in the short run, the period of interest to policymakers What is needed, as Leijonhufvud (1968) has argued so convincingly, is a short-run theory that can explain the dynamics of an economy in which the coordination of economic activities is not costless Only after this has been accomplished can we begin to understand the importance of Keynesian economics and its relation to the quantity theory. All this study shows is that the quantity theory can still be valid in the long run even in such a world of costly coordination.[36]

One of the lessons of Keynesian economics is that the short-run behavior of an economy is very much affected by the intermarket connections, other than price changes, that make one transactor's decisions dependent upon other transactors' expenditures In particular, this is the essence of the multiplier process Our problem is that the only short-run theory capable of describing such connections in a multimarket context is the textbook Keynesian theory which postulates that one set of prices (money wages) is completely inflexible On the other hand, the only theory that allows all prices to adjust is the Walrasian theory of Patinkin which does not allow for any of the intermarket connections because it cannot handle false trading [37] The difficulty that has stood in the way of allowing for false trading in a model like Patinkin's is that the microfoundations have not been strong enough to handle the problems mentioned above Because the "shopkeeper" approach used in this study attempts to deal with these problems it provides one way of allowing for the intermarket connections without postulating any absolute price rigidities [38] Thus it appears capable of handling some of the short-run questions that have traditionally been the exclusive domain of Keynesian economics

In fact, it can be shown[39] by using this approach that a monetary economy can behave in a Keynesian fashion in the short run even though it has a stable neoclassical long-run equilibrium In particular, there can be a Keynesian multiplier process at work in the short run that amplifies the effects of an exogenous change in aggregate demand Many practical macroeconomists have viewed Keynesian economics as important in the short run when prices have not fully adjusted to their equilibrium levels, and the quantity theory as important in the longer run when the economy approaches its long-run equilibrium [40] The present approach appears to provide a way of incorporating this view in a formal theory

References

Archibald, G C , and Lipsey, R G "Monetary and Value Theory A Critique of Lange and Patinkin " *Rev Econ Studies* 26 (October 1958) 1-22

Arrow, K J , Block, H D , and Hurwicz, L "On the Stability of the Competitive Equilibrium II " *Econometrica* 27 (January 1959) 82-109

---

[36] That is, costly in the sense that the activities of production, consumption and exchange cannot be suspended during the search for equilibrium prices

[37] See Leijonhufvud 1973

[38] See above

[39] Howitt 1973

[40] Of course this is not the view of the monetarists, who view the quantity theory as important in the short run also

Arrow, K J., and Hahn, F. H. *General Competitive Analysis*. San Francisco: Holden-Day, 1971.

Arrow, K. J , and Hurwicz, L. "On the Stability of the Competitive Equilibrium I." *Econometrica* 26 (October 1958) 522–52.

Barro, R J , and Grossman, H. I "A General Disequilibrium Model of Income and Employment." *A.E R* 61 (March 1971). 82–93.

Baumol, W J , Clower, R W , Burstein, M L , Hahn, F. H , Ball, R J.; Bodkin, R , Archibald, G C , and Lipsey, R. G "A Symposium on Monetary Theory " *Rev Econ Studies* 28 (October 1960) 29–56

Bushaw, D W , and Clower, R W *Introduction to Mathematical Economics*. Homewood, Ill. Irwin, 1957

Clower, R W "Permanent Income and Transitory Balances Hahn's Paradox " *Oxford Econ Papers* 15 (June 1963) 177–90

——— "The Keynesian Counterrevolution A Theoretical Appraisal " In *The Theory of Interest Rates*, edited by F H Hahn and F P R Brechling. London Macmillan, 1965

——— "A Reconsideration of the Microfoundations of Monetary Theory " *Western Econ J* 6 (December 1967) 1–9

——— "The Microfoundations of Monetary Theory Reply " *Western Econ. J.* 9 (September 1971) 304–5

Coddington, E A , and Levinson, N *Theory of Ordinary Differential Equations* New York McGraw-Hill, 1955

Edgeworth, F Y *Mathematical Psychics* London Kegan Paul, 1881

Fisher, F M "On Price Adjustment without an Auctioneer " *Rev. Econ Studies* 39 (January 1972) 1–15

Frevert, P "On the Stability of Full Employment Equilibrium " *Rev. Econ. Studies* 37 (April 1970) 239–51

Grossman, H I "Money, Interest, and Prices in Market Disequilibrium " *J P E* 79 (September/October 1971) 943–61.

Hadar, J "A Note on Stock-Flow Models of Consumer Behavior " *Q J E* 79 (May 1965) 304–9

Hahn, F H "Real Balances and Consumption " *Oxford Econ Papers* 14 (June 1962) 117–23

——— "On Some Problems of Proving the Existence of an Equilibrium in a Monetary Economy " In *The Theory of Interest Rates*, edited by F H Hahn and F P R Brechling London Macmillan, 1965

Hahn, F H , and Negishi, T "A Theorem on Non-Tâtonnement Stability " *Econometrica* 30 (July 1962) 463–69

Hicks, J R *Value and Capital* 2d ed Oxford Clarendon, 1946

Howitt, P W "Studies in the Theory of Monetary Dynamics " Doctoral dissertation, Northwestern Univ , 1973.

Kaldor, N "A Classificatory Note on the Determinateness of Equilibrium." *Rev Econ Studies* 1 (February 1934) · 122–36.

Leijonhufvud, A. *On Keynesian Economics and the Economics of Keynes*. New York Oxford, 1968

———. "Effective Demand Failures " *Swedish J Econ* 75 (March 1973). 27–48

Lloyd, C. "The Microfoundations of Monetary Theory: Comment." *Western Econ J* 9 (September 1971) · 299–303

Negishi, T "The Stability of a Competitive Economy · A Survey Article " *Econometrica* 30 (October 1962) 635–69.

Ostroy, J. M. "The Informational Efficiency of Monetary Exchange." *A.E.R*, 63 (September 1973) · 597–610.

Patinkin, D. *Money, Interest, and Prices* New York: Harper & Row. 1st ed., 1956; 2d ed., 1965.

Phelps, E. S. "Money Wage Dynamics and Labor Market Equilibrium." In *Microeconomic Foundations of Employment and Inflation Theory*, by E. S. Phelps et al. New York: Norton, 1970.

Samuelson, P. A. *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard Univ. Press, 1947.

Uzawa, H "The Stability of Dynamic Processes" *Econometrica* 29 (October 1961)· 617–31.

——— "On the Stability of Edgeworth's Barter Process." *Internat. Econ. Rev.* 3 (May 1962): 218–32.

Communications

The Interrelation between Labor and Capital Components of Racial Income Differences

George M. von Furstenberg

Indiana University

The isolation and projection of incomes from capital represents a recurring problem in the study of income inequality. Almost without exception, past studies of racial income or earnings differentials have attributed the residual remaining after skill adjustment to current discrimination or to other factors unconnected with property. If attempts to account for differences in capital holdings were made at all, they remained indirect and incomplete because of inadequate data.[1] To compensate some of this lack, the relation between the earnings, capital endowments, and incomes of the races is first clarified theoretically in this paper. The inference drawn from the model is then matched with what evidence there is on the actual racial distribution of incomes from capital.

It is shown that in distributional equilibrium with capital accumulation, the earnings of whites relative to blacks equal the ratio of their total incomes, provided the savings propensities, the rates of return on capital, and the growth rates of labor (population) are identical. This initial theorem and its less restricted forms yield equilibrium solutions of the ratios of capital per man between the groups. The mathematical relations involved can be useful to researchers dissatisfied with the statistical sources on the racial breakdown of capital ownership and income. They should also permit one to improve upon the rather indefinite treatment typically accorded to incomes from capital and property in intertemporal income comparisons for the two races. The two-sector model used in this paper is adapted from the pioneering work of Joseph Stiglitz (1969).

I am indebted to a referee and to Franz Gehrels, Michele Fratianni, Eric Herr, and Ralph Walter for helpful comments on earlier drafts.

[1] See, e.g., Bergmann 1971, p. 305.

The General Deduction

If y_i and w_i are the annual income and wage per capita of group i, c_i is the capital per man, and r_i is the rate of return on capital owned by group i rather than group j,

$$y_i = w_i + r_i c_i. \tag{1}$$

With savings per capita, s_i, assumed to be proportional to y_i, and with the savings propensity of group i denoted by m_i,

$$s_i = m_i y_i. \tag{2}$$

Since employment (and population) grow at the rate of n percent per year in group i, it follows from the definition of c_i that the time rate of change is

$$\dot{c}_i/c_i = s_i/c_i - n_i \tag{3}$$

Hence, the difference, D, in the rates of capital accumulation between groups i and j:

$$D = \dot{c}_i/c_i - \dot{c}_j/c_j = s_i/c_i - s_j/c_j + n_j - n_i \tag{4}$$

With the passage of time, capital per man would be equalized completely between the two groups if the initial inequality $c_i > c_j$ implied that D is negative (and vice versa) If w_i equals w_j and the population growth rates, r values, and savings propensities are also identical, as Stiglitz assumed, c_i/c_j must be greater than s_i/s_j as long as c_i exceeds c_j. Hence D is negative and c_j inevitably grows faster than c_i until it catches up with c_i However, equalization may remain incomplete in the racial context Substituting from (1) and (2) for s_i and s_j in equation (4) yields the general expression

$$D = m_i w_i/c_i - m_j w_j/c_j + r_i m_i - r_j m_j + n_j - n_i \tag{5}$$

In the special case where $m_i = m_j = m$, $r_i = r_j = r$, and $n_i = n_j = n$,

$$D = m(w_i/c_i - w_j/c_j). \tag{6}$$

If $(c_i/c_j) > (w_i/w_j)$, D remains negative as long as the degree of inequality of capital per man in groups i and j exceeds the corresponding inequality of the wage ratio Conversely, if capital endowments were more equal than the wage rates, c_i would grow faster than c_j until in equilibrium

$$(c_i/c_j)^* = w_i/w_j \tag{7}$$

Using (7), the ratio between the equilibrium levels of per capita income of groups i and j is also found to be equal to the wage ratio. For,

$$(y_i/y_j)^* = (w_i + rc_i^*)/(w_j + rc_j^*) = w_i/w_j. \tag{8}$$

8 (7 90) for the two races.[6] Hence, the passage of time alone should reduce the ratio of white to black capital incomes per person from 8 to between 4 and 2 The ratio would fall further to 1 7, the last term in (7a), if population growth rates as well as the rates of return on capital were equalized but the per capita earnings (and total incomes) of whites remained permanently 70 percent higher than those of blacks [7] As explained in the next section, a ratio of no more than 2 could also be the long-term growth as opposed to the instantaneous redistribution equilibrium value, even if the growth rates of population differ

Intertemporal Interpretation

From census data for families and unrelated individuals, the 1970 labor incomes per head were estimated as $w_{i0} = \bar{w}_i W_i/N_i$ and $w_{j0} = \bar{w}_j W_j/N_j$, yielding \$2,635 and \$1,550 for whites and blacks, respectively The corresponding capital incomes per head were \$300 and \$38 [8] Assuming the average rate of return on privately held capital assets was 5 percent *for both races,*[9] c_{i0} *equals \$6,000 and* c_{j0} *equals \$760* The weighted average of the two estimates of capital per head amounts to little more than half of \$10,000 This is to be expected since the census measure of capital incomes suffers from underreporting and does not include real capital gains, imputed rent, and returns on consumer durables.[10]

---

[6] In a related estimate standardized by age, Soltow (1972, p 83) reports wealth at age 50 of \$8,070 for all families and unrelated individuals in 1963, compared with \$1,180 for nonwhite households alone Though these figures are in 1870 dollars and include many more items such as household furnishings, the ratio is again over 7 A lower estimate of the difference is given for families in Kain and Quigley (1972, p 276) Home equities appear to be more equally distributed From the 1970 Census of Housing (1971b, pp 10, 41), current per capita home values of \$3,603 can be estimated for whites compared with \$1,205 for blacks Since the latter use mortgage financing to a lesser degree, the ratio of the much smaller home equities per head is probably less than 3

[7] This may be considered unlikely since the partial equalization of object capital held will presumably be paralleled by a reduction in the inequality of human capital accumulated Differences in human-capital endowments, however, which are frequently attributed to past discrimination, are one of the reasons for the wage differential between the races. Moreover, the crude labor participation rate, W_j/N_j, may itself be a function of racial income differences and birth rates

[8] Disregarding the approximately 2 million nonwhites not classified as blacks, the average level of capital income per head is therefore \$270 for all persons Multiplying by the 1970 population of 200 million yields \$54 billion This is less than one-half of \$114 billion, the sum of dividends, net interest, rental income of persons, and half of proprietors' incomes reported in the national income accounts for 1970 (By contrast, the wage and salary income of \$541 billion is underestimated by only 10 percent.) As on tax returns, underreporting of dividends and interest income seems to be widespread in the census surveys However, as long as the percentage of capital incomes (or assets) excluded is similar for blacks and whites, underreporting does not vitiate our estimates With regard to self-employment income, alternative ways of inferring the property share are discussed by Bronfenbrenner (1971, p 29)

[9] In the United States, the ratio of dividends to the price of stocks has been consistently just above 3 percent. Including capital gains raises the real return to about 4.5 percent per annum The sensitivity of results to variations in r is examined in n. 12.

To determine the speed with which the relative amount of white-owned capital per head may be expected to fall, the differential equation solutions are derived for equations (9) and (10). From equations (1)–(3) it is found that

$$\dot{c}_i = mw_i + (mr_i - n_i)c_i, \tag{9}$$

and

$$\dot{c}_j = mw_j + (mr_j - n_j)c_j. \tag{10}$$

Assuming the average productivity of labor, g, grows at the rate of 3 percent per annum and g equals the constant rate of growth in per capita real wages for whites and blacks, it follows that the ratio of per capita incomes at time t, with $r_i = r_j = r$, is given by

$$\begin{aligned} c_i(t)/c_j(t) = {} & \left\{e^{(mr - n_i)t}\left[c_{i0} + mw_{i0}\int_0^t e^{(g - mr + n_i)t}\,\mathrm{dt}\right]\right\} \\ & \div \left\{e^{(mr - n_j)t}\left[c_{j0} + mw_{j0}\int_0^t e^{(g - mr + n_j)t}\,\mathrm{dt}\right]\right\}. \end{aligned} \tag{11}$$

Substituting the initial values and the values of the parameters m, r, g, n_i, and n_j previously selected shows that the equilibrium ratio of c_i and c_j approaches the ratio of w_{i0} to w_{j0} as time goes to infinity. More precisely,

$$c_i(t)^*/c_j(t)^* = (g - mr + n_j)w_{i0}/(g - mr + n_i)w_{j0} = 2.0 \tag{12}$$

The larger g relative to n_j, the closer the approximation of the asset ratio to the ratio of labor earnings (1.7) for any given excess of n_j over n_i. This is a consequence of our assumption that as long as the aggregate rate of growth of population remains constant, g is invariant to differences in the growth rates of the white and black subgroups.[11] If the black population increases faster than the white, its share of total employment rises also. Hence, while the capital stock per man is diluted by faster relative rates of population growth within any separate group, labor income is not If mr is less than both n_i and n_j, so that net savings out of capital incomes alone do not suffice to sustain the initial level of capital per man, as our numbers imply, the cumulative savings out of labor incomes determine the ultimate ratios of capital per man. As time goes by, differences in population growth rates and in the initial capital endowments become unimportant. Even if $mr < n_j$ but $mr > n_i$, so that savings out of capital incomes more than suffice to maintain c_i but not c_j, savings out of labor incomes remain decisive.

---

[10] While imputed rent is at least included in the national income accounts, both construction and consumer durables were contained in the capital-stock measures previously cited

[11] To maintain the aggregate rate of population growth, both group rates must fall over time since the share of the faster-growing group in the total population rises. On the other hand, if the aggregate rate of growth of population were allowed to vary, the rate of growth of the average productivity of labor would be affected. For the last 2 decades, estimates of g range from 2.4 percent to 3.2 percent per annum. Fortunately, it is not necessary to be completely consistent in these respects, since changing the population growth rates has little effect on the ultimate equilibrium solution. Sensitivity tests for g are reported in the next note.

Assuming $c_i(t)/c_j(t)$ falls from its initial level of 8 to the long-run equilibrium value of 2.0, it is interesting to gauge how many years would have to pass for some specified part of the disequilibrium gap to be closed From equation (11) it is found that it would take 9 years to reduce the ratio from 8 to 4, and another 9 years to reduce it from 4 to 3[12] In all, it would take 50 years for the ratio of white to black capital per man to be brought down to 2 2 Looking backward, it is therefore entirely possible that the process of the partial equalization of capital holdings, begun by the Emancipation Proclamation[13] and reinforced in subsequent world wars, has still not run its course.

Since net asset data by race are incomplete for earlier periods,[14] it is not possible to test whether the convergence in capital holdings has actually continued during the last few decades at the rate suggested by the model. However, some indirect inferences can be drawn Extrapolating the numerator of equation (11) backward for 31 years ($t = -31$) yields a $c_i(t)$ of $2,315 for 1939 compared with $6,000 for 1970 The capital stock per black, $c_j(t)$, would fall even more. it becomes negative, indicating net indebtedness This result appears at variance with the findings of other researchers[15] Hence the previous parameters of the model appear to exaggerate the speed with which capital has been accumulated during the last 31 years Clearly the parameters chosen are not suitable for backcasting

---

[12] The results are not sensitive to the choice of parameters For instance, increasing the rate of return r from 5 to 10 percent without reducing c_{i0} or c_{j0} raises the time required to halve the inequality in capital per man from 9 to 9.4 years Reducing the rate of growth in productivity, g, from 3 percent to 2 6 percent has about the same effect

[13] At the end of the Civil War in 1865, "an estimated 80 percent of Negroes were totally illiterate and most were landless and destitute" (Kain 1969, p 7). Southern blacks, in particular, did not enjoy much security of property for decades thereafter, and significant net savings were unlikely to materialize under these conditions. For estimates of wealth for adult white and nonwhite males in 1870, see also Soltow (1972, p 83).

[14] Data on the racial distribution of some capital incomes (dividends, interest, net rental income, income from estates and trusts) were first reported in 1968, with net royalties added in 1970. The degree of instability shown by the figures for the first 2 years, which is characteristic of beginning data series, implies a caution against overinterpretation of the numerical estimates.

[15] Sterner (1943, p. 92) finds a positive surplus of income over expenditures for most classes of blacks even in 1935–36 (see also Lansing and Sonquist 1969, pp. 47–53). Significant net indebtedness, is, of course, unlikely to be obtainable for low-income households.

One possible reason for this is that the savings functions have negative intercepts. Then even with identical marginal propensities, the average savings propensity of blacks (m_j) falls short of the average propensity for whites (m_i), and the more so the greater the difference in incomes. For forecasting purposes m_j was therefore halved, and for backcasting it was reduced by 70 percent from its original level of 8.5 percent. With m_i of 0.085 and m_j of 0.0425 substituted for m in the numerator and denominator of equation (11), respectively, the equilibrium ratio of capital per man is found to rise from 2.0 to 4 2. Partly because the equilibrium level itself has risen, it now would take 30 rather than 6 years to reduce the ratio of capital holdings per man from 8 to 4 8 Similarly, backcasting with m_i of 0 085 and m_j of 0.025 implies that while whites owned an estimated $2,315 of capital per man in 1939, blacks owned $219, yielding a $c_i(t)/c_j(t)$ ratio of over 10.

Another reason for overestimating the speed with which capital holdings converge may be the disregarding of differences in the rates of return on capital between the races in the previous derivation. Even though Kain and Quigley (1972, p 270) reported that price appreciation of homes may be higher for blacks than for whites, the net return on home equities may still be lower if reduced benefits from leverage and tax deductions are taken into account[16] Furthermore, the alternative investments of most blacks, such as U S savings bonds and savings and loan shares and bank deposits, may be expected to yield a lower rate of return than the average return obtained by more diversified white investors To test for sensitivity to differences in r, r_j was therefore lowered from 5 percent to 3 percent. The equilibrium ratio of white to black asset holdings per head rose only from 2 0 to 2.1, but the convergence of capital holdings was retarded considerably[17] Once savings propensities are restored to equality, the capital holdings of blacks would again be implied to be negative in 1939. In other words, accumulation would still be sufficiently rapid to require negative wealth holdings in 1939 if blacks are to end up with no more than what they actually owned in 1970 under the assumed savings behavior. Hence varying r_j does not make backcasting any more feasible

Not surprisingly, the forecasting results were sensitive to *differences* in the average savings propensities assumed for whites and blacks However, because savings out of capital incomes are insufficient to keep up with population growth in this model, the sensitivity to differences between r_i and r_j was found to be smaller.[18] While it is interesting to let the average savings propensities and even the rates of return on capital coincide in long-term forecasting experiments, they must have differed systematically in the past, and they may still differ The reduction in the inequality of capital holdings per head would be slowed accordingly. Since the adjustment is gradual and since property incomes amount to a small fraction of total incomes, little immediate relief can be expected from the ongoing convergence of capital holdings Nonetheless, it is important, if not very comforting, to keep in mind that time will cure a significant part of one of the most glaring distributional inequities originally created by the condition of slavery All of our estimates suggest that this part is at least one-half, as when the ratio of asset holdings is reduced ultimately from 8 to 4, and it may be as high as three-quarters on the most optimistic assumption that all parameters (n, r, and m) other than labor incomes coverge for the two races.

---

[16] Kain and Quigley (1972, p 276) report that home equities amount to about 40 percent of the net worth of lower-middle-income white families and to 67 percent of the net worth of black families in the Survey of Economic Opportunity sample.

[17] Since the capital stock was inferred from capital incomes, as explained at the beginning of this section, c_{j0} rises from $760 to $1,267, yielding an initial asset ratio of below 4.8 for white to black. It would take 33 years to decrease this ratio by one-half. I owe this entire paragraph to the referee.

An Empirical Derivation of the Equilibrium Distribution of Assets

So far, all our results implied that the capital holdings of blacks are far from their distributional equilibrium. It is possible to test this inference empirically by asking what the capital endowment of blacks would have been if they held the same amount of assets in each income class as the U.S population as a whole This test may be ambiguous to the extent transition probabilities between income classes are higher for whites than for blacks On the other hand, except for recent immigrants, low-income whites may have been adversely selected from the generally higher-income white population Hence the comparison need not be biased · if low-income whites have larger negative transitory income components than blacks, they would be expected to own more assets in any income group However, whites would be expected to hold lower nonhuman wealth assets in equilibrium if poor whites are more likely to be the product of adverse self-selection than poor blacks,[19] who have fewer options to start with

On balance it is assumed in table 1 that blacks would have about the same amount of assets in any income class as the population as a whole if

---

[18] In long-term growth the sensitivity to differences in r is far less than in the instantaneous equilibrium discussed in the first sections of this paper because labor productivity is allowed to grow In the earlier model, equilibrium would have to be achieved through forced redistribution generating a current ratio of asset holdings that could be maintained without further redistribution at *given* labor incomes (and growth rates) for the races.

[19] Hanoch (1967) has suggested that private rates of return on higher education are lower for blacks than for whites For other references see Brimmer and Harper (1970, pp. 804–5). In a portfolio-balance model with imperfect substitutability between human and nonhuman capital, one would thus expect blacks to demand less human and more nonhuman capital than similarly situated whites if it is assumed that returns on nonhuman capital are identical for the two races. For a comparison of rates of return on human and nonhuman capital see Kendrick (1972, p 120).

TABLE 1

SIZE OF WEALTH BY INCOME CLASSES AND INFERRED EQUILIBRIUM WEALTH HOLDINGS FOR WHITES AND BLACKS, 1962

Household Income (*Y*) ($)	Average Wealth ($) (1)	*Y* Distribution White Households (%) (2)	Product (1 × 2) ($) (3)	*Y* Distribution Black Households (%) (4)	Product (1 × 4) ($) (5)
0–2,999	7,609	25 80	1,963	53 80	4,094
3,000–4,999	10,025	18 10	1,815	22 70	2,276
5,000–7,499	13,207	24 70	3,262	14 20	1,875
7,500–9,999	19,131	15 20	2,908	5 10	976
10,000–14,999	28,021	11 60	3,250	3 40	953
15,000–24,999	62,966	3 70	2,330	0 70	440
25,000–49,999	291,317	0 74	2,156	0 80	233
50,000–99,999	653,223	0 15	980	0 02	131
100,000 plus	1,698,021	0 01	170		
Total		100 000	18,834	100 00	10,978

SOURCES —Projector and Weiss (1966, p 110) U S , Department of Commerce, Bureau of the Census, "Income of Families and Persons in the United States 1962," *Consumer Income*, Series P-60, no 41 (Washington Government Printing Office, 1963), U S , Treasury Department, Internal Revenue Service, *Statistics of Income, 1962 Individual Income Tax Returns* (Washington Government Printing Office, 1965) (The latter source is used as a guide for distributing households with incomes above $25,000, the cutoff in the census source)

NOTE —In the source followed, the difference between the concept of wealth and that of net worth is small since wealth is reported net of the debts secured by specific assets Only unsecured personal indebtedness is not yet subtracted

equilibrium had already been reached in 1962 For that year, Projector and Weiss (1966) have provided estimates of the net wealth of households by income classes, which can be matched with the income distribution of black and white families and unrelated individuals. If only income matters and race plays no part, the predicted equilibrium ratio of capital per white over black household is $18,834/$10,978, or 1 7, while conversion to a per capita base yields $5,955/$3,090, or 1 9, for 1962

This result confirms dramatically that the existing inequality in wages or incomes implies a hypothetical equilibrium ratio in capital stocks per man of around 2 At least this would follow if blacks had asset holdings equal to those for the rest of the population with incomes distributed like those of blacks Since the actual asset ratio is probably close to 8 and almost certainly no less than 4, current income differences do not provide a sufficient explanation for differences in capital holdings between the races so that blacks are not just like poor whites Rather, reference must be made to the historical conditions and externally imposed handicaps specific to blacks if the continuing distributional disequilibrium in capital accumulation is to be accounted for

References

Bergmann, B. R. "The Effect on White Incomes of Discrimination in Employment." *J.P.E.* 79 (November/December 1971) 294–313.

Bishop, G. A. "Income Redistribution in the Framework of the National Income Accounts." *Nat. Tax J.* 19 (December 1966). 378–90.

Brimmer, A. F., and Harper, H. "Economists' Perception of Minority Economic Problems: A View of Emerging Literature." *J. Econ. Literature* 8 (September 1970): 783–806.

Bronfenbrenner, M. *Income Distribution Theory.* Chicago: Aldine, 1971.

Gaathon, A. L. "Economic Productivity in Israel, 1950–1965." *Rev. Income and Wealth* 16 (March 1970): 1–18.

Galenson, M "Do Blacks Save More?" *A.E.R* 62 (March 1972). 211–16.

Hanoch, G. "An Economic Analysis of Earnings and Schooling " *J. Human Resources* 2 (Fall 1967): 310–29

Kain, J F *Race and Poverty The Economics of Discrimination* Englewood Cliffs, N J . Prentice-Hall, 1969

Kain, J. F , and Quigley, J. M. "Housing Market Discrimination, Homeownership, and Savings Behavior." *A.E.R.* 62 (June 1972). 263–77.

Kendrick, J. W "The Treatment of Intangible Resources as Capital " *Rev. Income and Wealth* 18 (March 1972) 109–25.

Lansing, J B , and Sonquist, J "A Cohort Analysis of Changes in the Distribution of Wealth " Studies in Income and Wealth, vol. 33. *Six Papers on the Size Distribution of Wealth and Income.* Edited by L. Soltow New York Nat. Bur. Econ Res., 1969

Projector, D S , and Weiss, G. S. *Survey of Financial Characteristics of Consumers.* Washington Board of Governors, Federal Reserve System, 1966.

Soltow, L "A Century of Personal Wealth Accumulation " In *The Economics of Black America,* edited by H G. Vatter and T Palm. New York. Harcourt Brace Jovanovich, 1972

Sterner, R. *The Negro's Share* New York. Harper, 1943.

Stiglitz, J. E. "Distribution of Income and Wealth among Individuals." *Econometrica* 37 (July 1969) 382–97.

Tice, H S "Depreciation, Obsolescence, and the Measurement of the Aggregate Capital Stock of the United States, 1900–1962 " *Rev. Income and Wealth* 13 (July 1967) 119–54.

U S., Bureau of the Census "Income in 1970 of Families and Persons in the United States " *Current Population Reports.* Series P-60, no 80 Washington Government Printing Office, 1971. (*a*)

——— *1970 Census of Housing.* "General Housing Characteristics United States Summary " Washington Government Printing Office, 1971. (*b*)

An Optimal Program for Managing Common Property Resources with Congestion Externalities

Gardner Brown, Jr.

University of Washington

The purpose of this paper is to derive an optimum program for managing a common-property natural resource—fish, fowl, and some groundwater resources—whose rate of growth of the resource stock depends on the resource stock level and the current rate of "extraction." Looking forward in time, the dynamic property of this problem is provided by the fact that an instantaneous increase in the natural stock of a resource forever increases the physical productivity of that resource, while the externality aspect is of course inherent in the common-propertyness assumption. Congestion externalities arise from a specification of the aggregate production function which implies interdependent microproduction functions

It will be shown that if the resource stock initially is sufficiently small it will be socially optimal to have a period of no extraction When extraction begins, it and the variable factor increase faster than the resource stock is increasing Charges reflecting the imputed marginal social value of the resource are a decreasing function of the level of the stock These results differ from Vernon Smith's recent static treatment of the same problem (Smith 1968)[1] and are discussed below.

The Model

The rate of resource extractions, x, depends on the stock of that resource, X, and a variable factor, K:

$$x = G(X, K). \tag{1}$$

It is assumed that G is strictly concave and homogeneous of degree 1

The natural (percentage) rate of growth of the resource stock is denoted by

$$f(X), f' < 0, f'' \leq 0, f'(X^0) = 0, \tag{2}$$

I am indebted to my colleagues F. Mills, and Orris Herfindahl and to a referee for helpful comments, but am responsible for all errors

[1] But see Quirk and Smith (1969) for the correct formulation and solution of a similar problem. Arnold Zellner brought this paper to my attention after my unrevised version was completed.

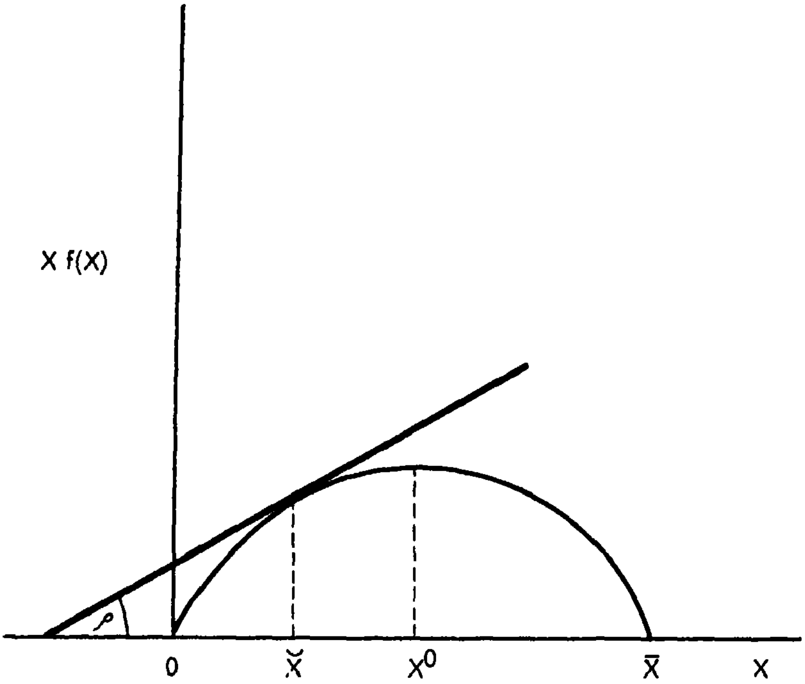


Fig 1

where "prime" refers to the derivative of f with respect to X In the case of the fishery, $0 \leq X \leq \bar{X}$, as illustrated in figure 1 For a rough approximation, fishery population-dynamics models express the population equation as $f(X) = r(1 - X), \bar{X}$, where r and $\bar{X}$ are parameters. Referred to as the Verhulst-Pearl logistic, it has been used for analyzing halibut (Crutchfield and Zellner 1962), yellowfin tuna (Schaefer 1967), and other species (Beverton and Holt 1957) An alternative form for $f(X)$ is the Beverton-Holt relationship $1/(\gamma + \beta X)$, where γ and β are parameters Originally this was applied to fish populations, but recently a slightly modified version seemed to provide an empirically adequate characterization for the North American mallard waterfowl population (Brown and Hammack 1971) Groundwater optimization models in practice generally assume that the natural absolute yield is independent of the stock level (Brown and McGuire 1967; Burt 1967) and therefore can be described by the Beverton-Holt relationship with $\beta = 0$. In fact, the geohydrology of aquifers is so varied that the signs of f' and f'' depend on the particular basin studied (Walton 1970)

It is assumed that the agency in control of the resource wishes to maximize the present discounted value of profit over time, and it pursues this objective assuming a constant product price, P, a constant variable factor price w, and a constant discount rate, ρ

$$\int_0^\infty [PG(X, K) - wK]e^{-\rho t}\, dt. \tag{3}$$

More generally, the net value function under the integral can be denoted by $V(x, X, K)$, where K can be defined in terms of boats in the case of the fishery, number and location of wells in the instance of groundwater, and hunters in the case of waterfowl. In these examples a good case for the existence of congestion externalities can be made The use of thinning and cutting crews and doses of fertilizer by managers of forests do not appear to cause externalities

From the homogeneity assumption, $G(X, K) = XG(1, k) \equiv Xg(k)$, where $k = K/X$. Rewriting (3) in terms of k, the agency now wishes to maximize

$$\int_0^\infty \{[Pg(k) - wk]X\}e^{-\rho t}\, dt, \tag{3'}$$

subject to

$$x \geq 0, \tag{4}$$

$$\dot{X} = Xf(X) - x, \tag{5}$$

$$x = Xg(k), \tag{6}$$

and given $X(0)$ This problem can be solved by the application of Pontryagin's Maximum Principle (Pontryagin et al 1962) The appropriate current value Hamiltonian (Arrow 1968, p 94) is

$$H = \{[Pg(k) - wk]X + \lambda[Xf(X) - Xg(k)]\}. \tag{7}$$

The auxiliary variable, λ, may be interpreted as the imputed marginal value of the stock resource

In addition to satisfying (4), (5) and (6), the necessary conditions for a maximum are

$$\frac{\partial H}{\partial K} = (P - \lambda)g' - w \leq 0, \tag{8}$$

$$\dot{\lambda} - \rho\lambda = -\frac{\partial H}{\partial X} = -[(P - \lambda)(g - kg') + \lambda(Xf' + f)], \tag{9}$$

and the transversality conditions,

$$\lim_{t\to\infty} X\lambda e^{-\rho t} = 0, \lim_{t\to\infty} \lambda e^{-\rho t} = 0. \tag{10}$$

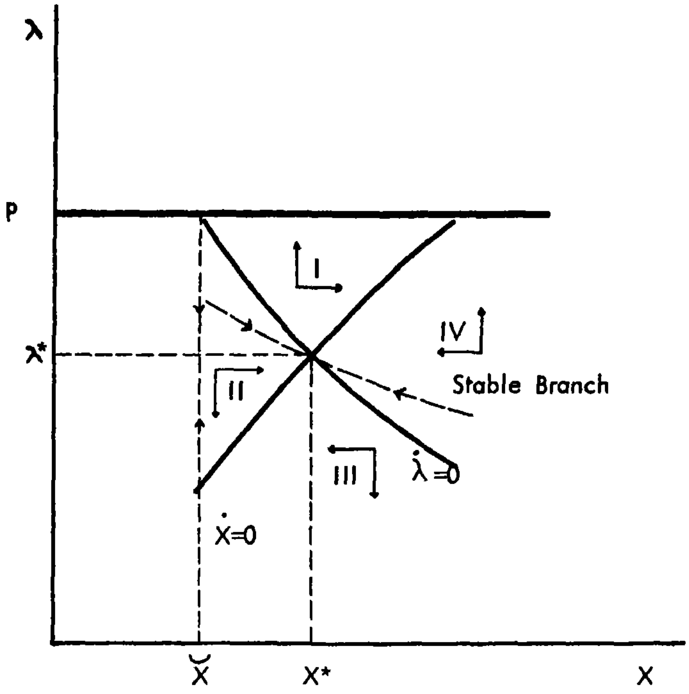


FIG 2

The sufficient conditions for an optimal policy established by Arrow (1968, pp. 91, 92, 95, 97) are satisfied in this problem essentially by assuming that the production function is concave, the conditions on natural resource growth specified by (2), and the assumption that $P > \lambda$.[2]

Equations (5), (6), (8), and (9) comprise an autonomous, two-differential-equation system in the sense that time is not an explicit variable (Arrow 1968, pp. 91, 95, 96, Arrow and Kurz 1970, pp 64–66). Thus, the optimal path is independent of initial values of X The system can be characterized in the following phase diagram (fig 2) The underlying mathematics are in the Appendix The motion of this system in phase space has the following economic interpretation A manager could establish a charge per unit of resource extracted sufficient to keep the stock level of the resource unchanged. If he levies a lower charge per unit extracted, harvest in excess of the natural amount of growth will occur, and the stock level will fall. This explains the leftward direction of the horizontal arrows in quadrants III and IV. Correspondingly, a charge too high reduces harvest below the physical equilibrium level, and the stock builds up as is indicated by the rightward horizontal arrows in quadrants I and II.

---

[2] Define $\bar{H}(X, \lambda, t) = \max_{k} H(X, k, \lambda, t)$. The assumption that $\bar{H}$ is concave in X for given λ and t implies that the necessary conditions are sufficient for optimality. It is easy to show that $H_{XX} = (P - \lambda)[(g''k^2)/X] + (2f' + Xf'') < 0$.

For a given λ there exists an X such that $\dot{\lambda} = 0$ Should the manager charge that same λ for a higher X, more extraction than is desired will take place (because cost is falling in the stock variable), the stock will be reduced below the desired levels, its natural marginal productivity will rise, and λ will rise to reflect this fact, expressed geometrically by the rising vertical arrows in quadrants I and IV. Falling vertical arrows in quadrants II and III follow from analagous reasoning.

It bears emphasizing that other phase diagrams can be drawn for this problem Clearly, there are natural resources for which it never is profitable to harvest because the unit extraction cost is higher than the price consumers are willing to pay Just as obvious, there can be a discount rate large enough, a biological growth rate low enough, for any level of X and a w small enough to warrant economically the extinction of a species (Quirk and Smith 1969, pp 20–21) The blue whale may be an example of the latter case if society believes the loss of this species is relatively low

Transient States

The natural place to begin the descriptive analysis is at low initial levels of the stock resource, $X(0) < X^*$, where the asterisk refers to the optimal stationary value This condition arises when valuable natural resources in the past have been regarded as common property by the users For X sufficiently small, its own marginal rate of return, $\partial \dot{X}/\partial X = f + Xf'$, is very high The marginal stock resource value, λ, will be greater than the market price of the harvestable product Therefore, from (7) and (8), the variable factor $K = 0$ and there should be no extraction Since it is specified that the marginal product of natural capital falls with increases in X, ultimately the optimal policy calls for harvesting, the inequality in (8) no longer is binding, and we follow the familiar rule that k should be so chosen that marginal cost, w/g' equals the net price of $P - \lambda$

From the individual user's point of view, when X is small and unmanaged, his marginal cost of extraction is high and he cannot fully capture in the future what he does not harvest today, so his marginal valuation of natural capital is low and rises as the stock of natural resources grows Of course the evolution of λ from the manager's point of view is just the reverse Initially the manager should set λ high because in future periods he can capture the high return from investing in the present As capital accumulates, investment in natural resources becomes less attractive relative to ρ, the rate of return available elsewhere Accordingly λ falls. The optimal path in the phase diagram simply corroborates good economic sense.

◆

As the resource grows, the rate of extraction, x, and the variable factor, K, grow over time until the optimal stationary level x^*, X^*, and K^* are reached, given that $x > 0$. Then the rate of harvest equals nature's periodic yield, $X^*f(X^*)$.

When $X < X^*$ and harvest is positive, both x and K increase faster than the stock resource is increasing. These conclusions follow from rewriting (6) as

$$\frac{x}{X} = g(k) = g\left(\frac{K}{X}\right) \tag{6'}$$

Since λ falls over time, g' must be decreasing in order for (8) to hold, so K must increase faster than X. Therefore, from (6'), output must be growing relatively faster than the stock resource.

The economic interpretation of these changes is that a rise in the net value of the harvested resource, $(P - \lambda)$, through a fall in λ induces an expansion of output and an increase in k, thereby increasing the marginal cost of output—through a decrease in the marginal productivity of k—in order to restore equality between net value and marginal cost. This particular qualitative relationship between optimal changes in x and K with respect to the state variable, X, is a result of the homogeneity assumption on the harvest production function. But for these conclusions, the content of the analysis throughout the paper is unchanged for any G which is concave.

In those fortunate instances when a manager comes upon a resource whose stock level is above the optimum, $X(0) > X^*$, the imputed price (λ) of the stock is initially relatively small and should grow over time along the optimum path, and the ratio of extraction to the stock level (x/X) falls over time since K/X is falling, until the steady-state level of the stock resource declines to X^*. This is illustrated by the right-hand portion of the stable branch in figure 2.

Stationary Values

It is of interest to inquire about the relationship between the steady-state stock level, X^*, and that level for which the annual physical yield of the resource is a maximum, X^0. In his static treatment of a comparable problem, Vernon Smith concluded that the optimal stock level is always to the right of X^0 (see fig. 1) because the marginal extraction cost is falling in X (Smith 1968, p. 427). In the dynamic context, such a conclusion cannot be reached because there is an opportunity cost involved in waiting for larger stock levels. We know that $X^* > \breve{X}$ where $f(\breve{X}) + \breve{X}f'(\breve{X}) = \rho$, but this fact exhausts our information about the location of X^*.[3]

---

[3] From (9), $\dot{\lambda} = -(P - \lambda)(g - kg')$ when $f(\breve{X}) + \breve{X}f'(\breve{X}) = \rho$. Either $\lambda \geq P$, in which case X obviously is growing since there is no harvest, or $\lambda < P$, in which case $\lambda < 0$ since $g - kg' > 0$ because g is strictly concave.

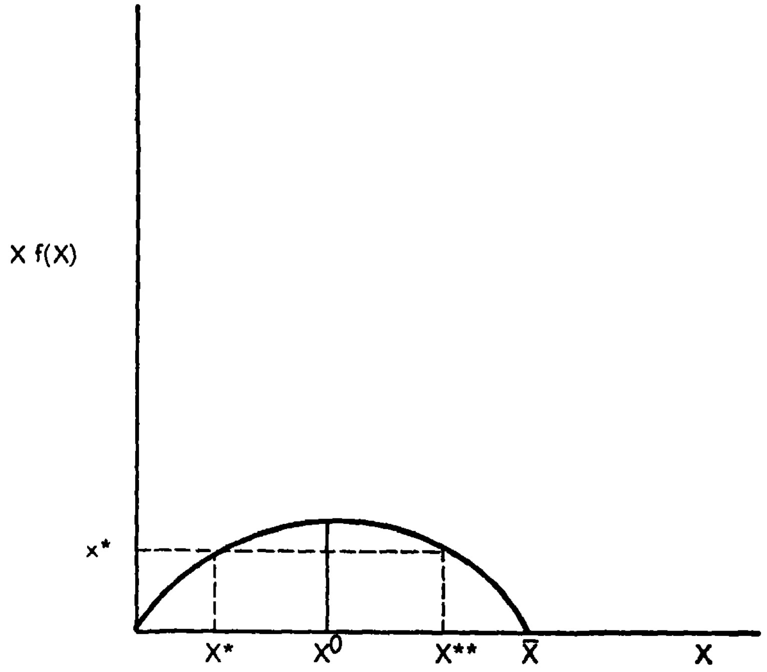


FIG 3

Although it is true that extraction costs fall with increasing X, they may not be falling sufficiently fast to warrant waiting while stocks build up

If $X^* < X^0 < X(0)$, the optimal management plan calls for progressively higher unit costs of extraction over time as X decreases Referring to figure 3, had the adjustment process toward a steady-state value of X^* begun at $\bar{X}$, it should not stop at X^{**}, where the instantaneous rate of profit is greater than at X^*, but it should go to X^* The "short-run" profit obtained from driving the stock level from X^{**} to X^* more than compensates for the "long-run" additional costs of extracting $f(X^*)$

This solution may be compared with the results for a related problem obtained by C. G. Plourde (1960) In his model, the utility of consumption $U(C)$, or utility of harvest, is maximized over time in a context of no cooperating inputs Plourde shows that the steady-state optimum natural resource stock level is that for which the own rate of return equals ρ Thus his stationary level of stock always is lower than that prescribed in this model The steady-state harvest also is lower compared with the levels obtained in the present analysis provided $X^* < X^0$. That is, the optimum stock is not to the right of the level which produces maximum physical sustained yield. In general, optimal harvest can be greater than, equal to, or less than the solution value satisfying Plourde's model Only by placing further restrictions on the functions can we make the comparative results unique.

The policy prescriptions of the two models also differ when natural resource stocks are comparatively low Since $U'(C) \rightarrow \infty$ as $C \rightarrow 0$, there should always be some consumption in Plourde's model, whereas in this study the imputed marginal value of the stock resource for small X more than matches the instantaneous rate of profit for any positive rate of extraction, so harvesting should be postponed Apart from this minor distinction, the behavior of the two systems in terms of $\dot{\lambda}/\lambda$ and $\dot{X}/X$ during the transient states are qualitatively equivalent

Resource Management Policy

Efficient natural resource use can be achieved within a decentralized competitive system if the management agency enunciates two pricing policies, one which reflects the scarcity value of the resource stock and another one designed to capture the opportunity cost of congestion Given the common-property aspects of this problem, it is necessary to levy a charge of λ per unit extracted, since λ is not market phenomena

The congestion externality may be accounted for by charging a tax for the use of a unit of the variable factor The genesis of the congestion externality arises from the fact that variable factors in a competitive economy will be committed to the harvesting activity until

$$(P - \lambda)\frac{Xg(k)}{k} = w, \tag{11}$$

since average output per variable factor in the resource industry is the firm's marginal output as viewed by it when the number of firms harvesting is large [4] To induce the correct level of K, a unit charge, t,

$$t = (P - \lambda)\frac{(g - g'k)}{k}, \tag{12}$$

on the variable factor is computed from the requirement of equality of (8) and (11) [5]

I am unaware of any domestic natural resource management program remotely resembling the approach recommended here Public natural resource managers in this country use neither a charge which varies with rate of use nor a charge for the right of entry, nor do they have the necessary enabling legislation to restrict the number of firms entering the fishery or groundwater basins. Additionally, the number of waterfowl hunters is not restricted[6] There are two exceptions to these statements In Colorado and New Mexico, application for a new right to use groundwater can be denied if it impairs unreasonably existing rights in terms of quantity or quality[7] Only in California can a public water district charge a pumping tax, and only one district makes use of this taxing power.[8]

---

[4] The condition that there are a large number of firms probably is met adequately for natural resources such as the various domestic fisheries and groundwater There may be some doubt about its validity in the instance of forestry and in some international fisheries such as whaling

[5] The tax is a simple expression if the extraction production function is $x = K^{\alpha}X^{1-\alpha}$, for then $g = k^{\alpha}$ and from (12) and (8), $t = w(1 - \alpha)/\alpha$ The variable factor tax is therefore independent of X and K.

Conclusion

The static treatment and solution of an externality problem which is inherently dynamic generally will lead to prescriptions which cause social losses This is illustrated in Vernon Smith's article (1968, p 426) where his artificial restriction of $\dot{X} = 0$ is binding in the dynamic formulation, hence costly

Omitted from this analysis are costs of exit in the case of overexploited stocks If, for example, the social cost of moving boats out of the fishery or of displacing loggers in the forestry industry are greater than the profit obtained from pursuing the policy described above, obviously the best social policy is to do nothing in the context of the present model Given exit costs smaller than this magnitude and assuming income transfers are politically inappropriate, the "second-best" policy is apparent although not formally developed Some absolute rate of extraction over time greater than that derived earlier is called for until the steady state is reached, and this happy event will be postponed in time

Appendix

Differentiating both (8) and (5) after substitution of (6) into (5) yields, after substitution,

$$\left.\frac{d\lambda}{dX}\right|_{X=0} = \frac{(P - \lambda)g''f'}{(g')^2} > 0, P > \lambda,$$

---

[6] No one can seriously argue that the cost of a season duck stamp is set with an eye to managing hunter congestion There are restrictions on the right of entry to shoot game in the public forests of Germany Curiously, the restriction is based on the ranking of individuals in terms of their performance on an exam which test hunting proficiency and knowledge of "good" conservation practices (see Webb 1961) As for international resources, the right to harvest pelagic seals is restricted in some instances (Kasahara 1972, p 66) Entry of Japanese fishermen is restricted by the central government or fishermen cooperatives in the inshore, offshore and high seas fisheries (ibid., chap. 1, and pp 110–11)

[7] See Fundingsland v Colorado Groundwater Commission, 468 P 2d 835 (1970), Mathers v Texaco, Inc. 77 New Mexico 239, 421 P 2d 771 (1966) For further discussion see Corker 1971, chap. 5

[8] Interview with Langdon Owen, manager, Orange County Water District, December 28, 1970.

•

since $f' \leq 0, g'' < 0$;

$$\left.\frac{d\lambda}{dX}\right|_{X=0} = \frac{(P-\lambda)g''f'}{(g')^2}, \text{ for } X > 0.$$

Differentiating both (8) and (9) and simplifying through application of (8) gives

$$\left.\frac{d\lambda}{dX}\right|_{\lambda=0} = \frac{\lambda(2f' + Xf'')}{\rho - (f + Xf') + g} < 0, P > \lambda > 0.$$

The numerator is negative because $f' < 0$ and $f'' \leq 0$. The denominator is positive because (9) can be rewritten as

$$\rho - (Xf' + f) = \frac{(P-\lambda)}{\lambda}(g - kg') > 0,$$

since strict concavity of g makes $g - kg' > 0$

From (10) and (8) it is easy to show that, when $\lambda = 0$,

$$\left.\frac{d\lambda}{dX}\right|_{\lambda\ \text{constant}} = -\lambda(2f' + Xf'') > 0, \tag{A1}$$

and from (5) and (8), it is readily proved that, when $X = 0$,

$$\left.\frac{dX}{d\lambda}\right|_{\lambda\ \text{constant}} = \frac{-(g')^2}{(P-\lambda)g''} > 0, P > \lambda \tag{A2}$$

The motion of this system is characterized by (A1) and (A2)

References

Arrow, K J "Applications of Control Theory to Economic Growth " Stanford, Calif Stanford Univ Inst Math Studies Soc Sci, 1968

Arrow, K J, and Kurz, Mordecai *Public Investment, the Rate of Return and Optimal Fiscal Policy* Baltimore Johns Hopkins Press, 1970

Beverton, R J H, and Holt, S J *On the Dynamics of Exploited Fishery Populations* London Her Majesty's Stationery Office, 1957

Brown, Gardner, Jr, and Hammack, Judd "Economic Evaluation of a Fugitive Recreational Resource Migratory Waterfowl " Paper prepared for Resources for the Future Multi-disciplinary Workshop on Research in Wildlands, Wildlife and Science Resources, Missoula, Montana, August 4–5, 1971

Brown, Gardner, Jr, and McGuire, C B "A Socially Optimum Pricing Policy for a Public Water Agency." *Water Resources Res* 3 (1967)· 33–43

Burt, Oscar "Groundwater Management under Quadratic Criterion Functions." *Water Resources Res* 3 (1967) 673–82

Corker, Charles "Groundwater Law, Management and Administration." Review draft of Legal Study 6, prepared for National Water Commission, October 1971

Crutchfield, James, and Zellner, Arnold "Economic Aspects of the Pacific Halibut Fishery." *Fishery Ind. Res.* 1 (April 1962) 1–173

Kasahara, Hiroshi "Future of Japanese Distant-Water Fisheries " Mimeographed. University of Washington, 1972

Plourde, C G. "A Simple Model of Replenishable Natural Resource Exploitation." *American Econ Rev* 60 (June 1960). 518–22.

Pontryagin, L., Boltyanskii, V.; Gamkrelidge, R.; and Mischenko, E. *The Mathematical Theory of Optimal Processes.* New York Interscience, 1962.

Quirk, James, and Smith, Vernon. "Dynamic Economic Models of Fishing." Research Paper no 22, Research Papers in Theoretical and Applied Economics, Univ. Kansas, Dept. Econ., June 1969.

Schaefer, Milner, "Fishery Dynamics and Present Status of the Yellowfin Tuna Population of the Eastern Pacific Ocean." *Inter-American Tropical Tuna Commission* 12, no. 3 (1967): 89–136.

Smith, Vernon. "Economics of Production from Natural Resources" *American Econ Rev* 58 (June 1968) 409–31.

Walton, William. *Groundwater Resource Evaluation.* New York: McGraw-Hill, 1970

Webb, William "Forest Wildlife Management in Germany." *J. Wildlife Management* (April 1961), pp 147–61

Cognitive Range in the Theory of Revealed Preference

Anthony Y. C Koo

Michigan State University

Peter Schmidt

University of North Carolina at Chapel Hill

In some recent work (Koo 1963, Koo and Hasenkamp 1972), empirical investigation has been made of the extent to which the behavior of households is in accordance with the axioms of revealed-preference theory These investigations have found that, with few exceptions, almost all families made at least some inconsistent choices Of course, this is not surprising to economists, and the literature abounds with explanations However, few of these explanations are testable against market data. In this paper we attempt to test two possible explanations for observed inconsistent behavior The hypotheses to be tested are proposed in Section I, and the tests are described in Sections II and III

I

Let $\bar{q}^i$ and $\bar{p}^i$ represent n-dimensional vectors representing a "bundle" of goods and their prices, respectively, and let $\bar{p}^i\bar{q}^i$ be the vector product representing expenditure on bundle $\bar{q}^i$ when prices are $\bar{p}^i$ Then the basic postulate of revealed preference states that if bundle $\bar{q}^i$ is chosen when prices are $\bar{p}^i$, and bundle $\bar{q}^j$ is chosen when prices are $\bar{p}^j$,

$$\bar{p}^i\bar{q}^i > \bar{p}^i\bar{q}^j \tag{1}$$

and

$$\bar{p}^j\bar{q}^j < \bar{p}^j\bar{q}^i, \tag{2}$$

then the choice of $\bar{q}^i$ is said to be a "consistent" choice On the other hand, the choice of $\bar{q}^i$ and $\bar{q}^j$ is "inconsistent" if

$$\bar{p}^i\bar{q}^i > \bar{p}^i\bar{q}^j \tag{3}$$

and

$$\bar{p}^j\bar{q}^j > \bar{p}^j\bar{q}^i \tag{4}$$

The authors wish to acknowledge with gratitude the services of Art Havenner in supervising the programming and computation of the project.

both hold. The basic postulate of revealed-preference theory is therefore just that such inconsistencies do not occur.

Two hypotheses can immediately be suggested in an attempt to explain the inconsistencies which are observed to occur. The first is that consumers cannot rank in a consistent manner alternatives which are very similar For example, a consumer may not be able to decide if he prefers 2 oz. of beer and 2 oz of wine, or 2.001 oz of beer and 1.999 oz of wine The assumption here is that an individual may only have discrete levels of perception (or preference), and that alternatives can be ranked consistently only if they differ by some minimum "threshhold" amount This hypothesis, which will be called the "threshhold hypothesis," is very similar to the threshhold notion as used in sensory psychology.

The second hypothesis is that consumers cannot rank in a consistent manner alternatives which are very dissimilar A consumer may have a "typical" consumption pattern and may not adequately consider alternative consumption patterns which differ widely from the one he is used to Hence, he would not be likely to make errors when he is ordering bundles in some immediate neighborhood of his past experience; he is likely to make errors in ordering bundles which are some distance apart

The assumption is equivalent to the definition of marginal rate of substitution between various goods as describing the individual's scale of preference, and the errors in ordering imply that the differential equation for indifferent changes in purchases may not be integrable [1] This hypothesis will be called the "cognitive-range hypothesis," since it essentially implies that a consumer has a limited cognitive range [2]

These two hypotheses will now be tested against market data of the Michigan State University consumer food panel consisting of a sample of about 250 families In the present study, only families that made all 52 weekly reports for at least 1 calendar year in 1955–58 were chosen, the number of families considered was 87 in 1955 and 1956, and 127 in 1957 and 1958 The 52 weekly purchases were grouped into 13 observations. Details of the data are described in Quackenbush and Shaffer (1960).

---

[1] For an early brief introduction to the subject of integrability, see Allen (1938, pp 438–42, 509–17) In economic terms why a succession of infinitesimal indifferent displacements could not lead to an indifference variety of some sort rests in the absence of transitivity (see Georgescu-Roegen [1967, pp 149–50 and sec IV], Hurwicz [1971], and Samuelson [1950] That transitivity in the revealed-preference sense also requires integrability was first shown by Houthakker (1950, pp 173–74)

[2] One may argue that the cognitive-range hypothesis is really testing the income effects on consumer behavior When his income remains unchanged, the consumer can only substitute one bundle of goods for another It follows that he is less likely to make a big adjustment and will, accordingly, commit fewer mistakes On the other hand, a large adjustment of his consumption pattern frequently follows from an income change, in which case he is more likely to make mistakes This income variation hypothesis was tested in Koo and Hasenkamp (1972) and rejected. The number of inconsistent bundles for each family was regressed against family characteristics including annual family income We have a separate equation for each year (1955, 1956, 1957, and 1958) based on cross-sectional data. The signs for the coefficients of family annual income are equally divided between plus and minus for the four equations.

II

Consider two bundles $\bar{q}^i$ and $\bar{q}^j$, and suppose that $\bar{p}^i\bar{q}^i > \bar{p}^i\bar{q}^j$ Then the choice between $\bar{q}^i$ and $\bar{q}^j$ is consistent if $\bar{p}^j\bar{q}^j < \bar{p}^j\bar{q}^i$, and is inconsistent if $\bar{p}^j\bar{q}^j > \bar{p}^j\bar{q}^i$ Now, the threshhold hypothesis could be interpreted as implying that inconsistencies would be more likely when $|p^jq^j - p^jq^i|$ is small than when it is large—that is, inconsistencies might occur when $\bar{p}^j\bar{q}^j$ and $\bar{p}^j\bar{q}^i$ are almost imperceptibly different Note that this can be interpreted as a threshhold effect in the perception of expenditure levels, representing an inability or unwillingness to make sufficiently fine expenditure calculations Alternately, since $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i| = |\bar{p}^j(\bar{q}^j - \bar{q}^i)|$, it can be interpreted as a threshhold effect in the preference ordering of various commodity bundles, if $|\bar{p}^j(\bar{q}^i - \bar{q}^j)|$ is interpreted as a measure of the "difference" between $\bar{q}^i$ and $\bar{q}^j$ In either case, the conjecture is that $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i|$ should tend to be smaller for inconsistent choices than for consistent ones

To test this conjecture, the following procedure was employed For each family year, $\bar{p}^i\bar{q}^i$ was compared with $\bar{p}^i\bar{q}^j$ for all $i \neq j$, $i, j = 1, 2, \quad 13$ If $\bar{p}^i\bar{q}^i < \bar{p}^i\bar{q}^j$, nothing more is done On the other hand, if $\bar{p}^i\bar{q}^i > \bar{p}^i\bar{q}^j$ (so that $\bar{q}^i$ is revealed preferred to $\bar{q}^j$), the observation is placed in one of two categories

If $\bar{p}^j\bar{q}^j < \bar{p}^j\bar{q}^i$, the choice is consistent and the difference $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i| = \bar{p}^j\bar{q}^i - \bar{p}^j\bar{q}^j$ is entered into the consistent list

If $\bar{p}^j\bar{q}^j > \bar{p}^j\bar{q}^i$, the choice is inconsistent and the difference $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i| = \bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i$ is entered into the inconsistent list

We then tested to see whether the mean of the entries in the consistent list equaled the mean of the entries in the inconsistent list, the above conjecture is, of course, that the mean of the entries should be larger in the consistent list This test is carried out using the test statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{(Sp^2/n_1) + (Sp^2/n_2)}, \text{ where } Sp^2 = \frac{n_1S_1^2 + n_2S_2^2}{n_1 + n_2 - 2},$$

and where n_1, $\bar{X}_1$, and S_1^2 are the number of entries, estimated mean, and estimated variance for the consistent list, n_2, $\bar{X}_2$, and S_2^2 correspond to the inconsistent list The relevant distribution (subject to the usual conditions) is the t-distribution with $n_1 + n_2 - 2$ df Clearly, a positive test statistic is needed to confirm the conjecture made above.

One such test was performed for each family year, and the results were rather mixed Perhaps the best method of summarizing the results is just to compare the number of positive statistics to the number of negative statistics. This is done in table 1. Of 422 family years for which there were

TABLE 1
PAIRWISE COMPARISONS—ABSOLUTE DIFFERENCES

Year	Positive Statistics*	Negative Statistics†	No Statistics‡
1955	29	52	3
1956	27	51	6
1957	30	92	5
1958	43	78	6
Sum	129	273	20

* Indicates larger mean for consistent group than for inconsistent group
† Indicates smaller mean for consistent group than for inconsistent group
‡ Indicates no observations in either the consistent or inconsistent group

TABLE 2
PAIRWISE COMPARISONS—PERCENTAGE DIFFERENCES

Year	Positive Statistics	Negative Statistics	No Statistics
1955	24	58	3
1956	25	55	6
1957	29	95	5
1958	36	79	7
Sum	114	287	21

observations in both the consistent and inconsistent groups, there were 129 positive statistics and 273 negative statistics In other words, the average entry actually tended to be larger in the inconsistent group than in the consistent group, contrary to the conjecture based on the threshhold hypothesis Indeed, if we test the null hypothesis that positive and negative statistics are equally likely, it is rejected at any reasonable level—the reason being that under the null hypothesis the number of negative statistics has mean 201 and standard deviation 10 025, and the observed number (273) is over 7 standard deviations away from 201 Hence, the preponderence of negative statistics is statistically significant

It is therefore apparently true that $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i|$ is typically greater for inconsistent choices than for consistent ones This result can be interpreted as supporting the cognitive-range hypothesis and contradicting the threshhold hypothesis

As a further check, the same procedure was followed except that in this case we used percentage differences

$$\frac{|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i|}{\frac{1}{2}(\bar{p}^j\bar{q}^j + \bar{p}^j\bar{q}^i)},$$

rather than just the absolute differences $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i|$ The results are given in table 2 and are very similar to those reported in table 1

TABLE 3
PAIRWISE COMPARISONS—ABSOLUTE DIFFERENCES

Year	Positive Statistics	Negative Statistics	No Statistics
1955	34	46	2
1956	34	43	6
1957	50	73	4
1958	55	66	7
Sum	173	228	19

III

In the tests above, we considered bundles $\bar{q}^i$ and $\bar{q}^j$ such that $\bar{p}^i\bar{q}^i > \bar{p}^i\bar{q}^j$, and then looked at the difference (in absolute or percentage terms) between $\bar{p}^j\bar{q}^j$ and $\bar{p}^j\bar{q}^i$ It might be objected that the difference between $\bar{p}^i\bar{q}^i$ and $\bar{p}^i\bar{q}^j$ is also relevant, since it, too, can be used to measure the "difference" between $\bar{q}^i$ and $\bar{q}^j$ Therefore, the first test reported in Section II was rerun, using the smaller of $|\bar{p}^i\bar{q}^i - \bar{p}^i\bar{q}^j|$ and $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i|$ rather than just $|\bar{p}^j\bar{q}^j - \bar{p}^j\bar{q}^i|$ The results are given in table 3 Once again, the average difference is typically greater for the inconsistent group

The differences which are being used in these tests can be illustrated graphically in the two-goods case (X_1 and X_2) Suppose that a consumer purchases three bundles of goods (at three different points in time), $\bar{q}^0$ at price $\bar{p}^0$, $\bar{q}^1$ at price $\bar{p}^1$, and $\bar{q}^2$ at price $\bar{p}^2$ The three bundles are represented by one point each in figure 1 To avoid cluttering the diagram, we let the price lines $\bar{p}^0$ and $\bar{p}^2$ coincide in the diagram

It is easily seen that $\bar{q}^0$ is revealed preferred to $\bar{q}^1$ and that the choice is consistent Now, if we draw a (dotted) line through $\bar{q}^0$, parallel to $\bar{p}^1$, *then the difference* $|\bar{p}^1\bar{q}^1 - \bar{p}^1\bar{q}^0|$ *can be seen (in terms of* X_2*) as the* radius of the smaller circle on the vertical axis, or (in terms of X_1) as the radius of the smaller circle on the horizontal axis Similarly, the choice between $\bar{q}^1$ and $\bar{q}^2$ is inconsistent, and the difference $|\bar{p}^1\bar{q}^1 - \bar{p}^1\bar{q}^2|$ can be seen as the radius of the larger circle on the vertical axis or the larger circle on the horizontal axis Note that the figure is constructed so as to support the cognitive-range hypothesis

In conclusion, our evidence seems to support the hypothesis that the cognitive range of consumers is limited, and that the consumer seems to be able to better discriminate between small changes from his established purchases than between widely different sets of purchases The finding thus points to the empirical relevance of theoretical work on integrability conditions

References

Allen, R G. D *Mathematical Analysis for Economists.* London · Macmillan, 1938.
Georgescu-Roegen, N. "The Pure Theory of Consumer's Behavior." *Q J.E.* 50

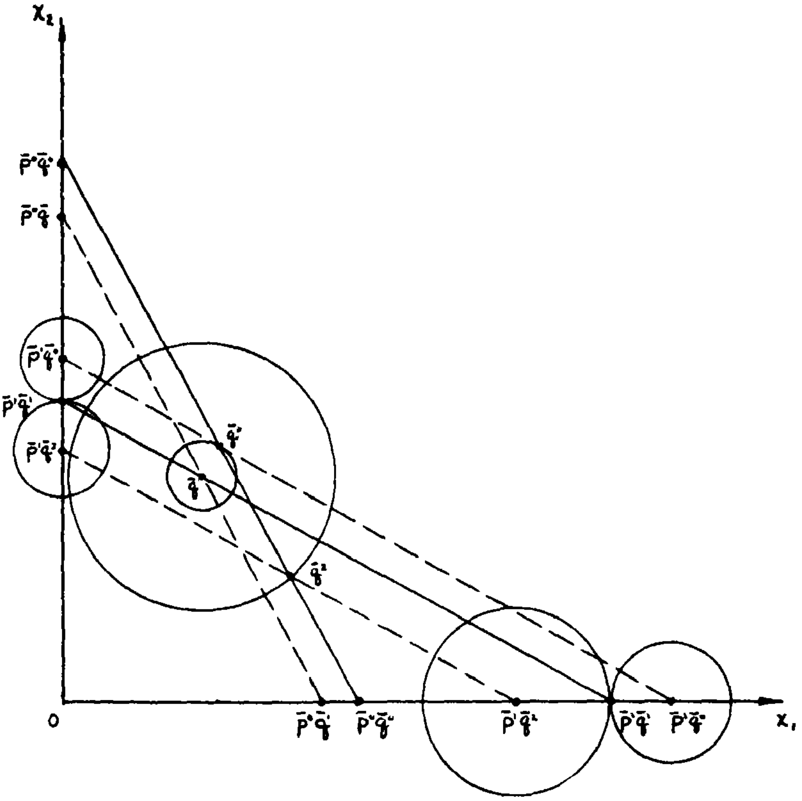


Fig 1

(August 1936) 545–93, reprinted in *Analytical Economics Issues and Problems* Cambridge, Mass Harvard University Press, 1967

Houthakker, H S "Revealed Preference and Utility Functions " *Economica*, n s, 17, no. 66 (May 1950) 159–74

Hurwicz, L. "On the Problem of Integrability of Demand Functions," in *Preference, Utility and Demand*, edited by J S Chipman, L Hurwicz, M K Richter, and H F. Sounenschen New York Harcourt Brace Jovanovich, 1971

Koo, A Y. C "An Empirical Test of Revealed Preference Theory " *Econometrica* 31, no 4 (October 1963) 646–64

Koo, A Y C., and Hasenkamp, G "Structure of Revealed Preference—Some Preliminary Evidence " *J P E* 80, no 4 (July-August 1972) 724–44.

Quackenbush, G. G., and Shaffer, J D *Collecting Food Purchase Data by Consumer Panel*. Agricultural Experiment Station, Technical Bull 279, Michigan State Univ, 1960.

Samuelson, P. A. "The Problem of Integrability in Utility Theory." *Economica* 17, no. 68 (November 1950): 355–85; reprinted in *The Collected Scientific Papers of Paul A. Samuelson*, edited by J E. Stiglitz Cambridge, Mass.: M.I T Press, 1966.

◆

Consumers' Durable Goods, Human Capital, and the Pareto Optimality of the Lange-Lerner State

Michael R Darby

University of California, Los Angeles

This note shows that, in the presence of consumers' durable goods or human capital, the Lange-Lerner state fails to achieve Pareto optimality even in the limited sense of the current period's consumption and production This follows because different prices exist in the same market, and this difference leads to distortions in the production-consumption decision, entirely apart from the question of optimal intertemporal allocation

It is well known that Oskar Lange and Abba Lerner developed an economic theory of socialism in which civil servants operate firms as if profit maximizing in competitive factor and product markets It has been generally accepted that, under the assumptions of the model and aside from certain highly special conditions, static Pareto optimality in the current period's production and consumption will obtain, so that the discussion has focused on other points [1] This assertion is easily shown to be in error, however, if it is recalled that some of the goods being purchased by the household are consumers' durable goods

In Lerner's *The Economics of Control,* consumers were allowed to participate in the government capital market by borrowing or lending at the same rate of interest as applied to firms [2] This has generally been taken to be inconsistent with socialist ideology, so I will be considering the dominant assumption of Lange in which private lending and borrowing are outlawed and the interest on savings deposits at the government bank is arbitrary and "would not have any necessary connection with the marginal net productivity of capital" (Lange 1938, pp 85–86) That is, the government firms bid in the capital market for the arbitrarily fixed

---

The author would like to acknowledge helpful comments from two anonymous referees which have significantly improved the exposition

[1] See, for example, I M D Little (1957, pp 260–66) and F M Bator (1958, pp. 353–56) A recent article by P. C. Roberts (1971) contains a survey of the literature.

[2] However, there were to be restrictions on total amounts (Lerner 1946, p. 269), so for those individuals at the maximum, the following discussion applies.

(by the central planning board) amount of capital, thus determining the interest rate, r. The interest rate on savings is assumed to be $r_s < r$[3]

Consider the purchase and rental prices of a consumers' durable good under the socialist regime Take X_1 and p_1 as the quantity and price of a pure service which enters directly into the utility function or can be combined with the consumers' durable good, say X_3 at a price p_3, to produce a flow of services, X_2 Assume that p_3 correctly reflects the social marginal cost[4] The production function of X_2 is

$$X_2 = f(X_1, X_3) \tag{1}$$

Now the cost function of a firm producing X_2 is

$$C(X_2) = p_1X_1 + (r + \delta)p_3X_3, \tag{2}$$

where the durable good depreciates (by evaporation) at the rate δ per period The profit function to be maximized is

$$\pi = p_2 \ f(X_1, X_3) - p_1X_1 - (r + \delta)p_3X_3 \tag{3}$$

The first-order conditions for a maximum require that

$$\frac{\partial\pi}{\partial X_1} = p_2\frac{\partial f}{\partial X_1} - p_1 = 0, \tag{4}$$

$$\frac{\partial\pi}{\partial X_3} = p_2\frac{\partial f}{\partial X_3} - (r + \delta)p_3 = 0 \tag{5}$$

Therefore, the service flow X_2 will be priced such that

$$p_2 = \frac{r + \delta}{\partial f/\partial X_3} \ p_3 \tag{6}$$

Now broaden the consumers' choice to include not only X_1 and X_2 but also X_3 Consider the consumer who can save or borrow at the state bank and thus has an internal rate of discount $= r_s$[5] If he were to reduce his savings so as to purchase some X_3, he could produce X_2 for himself at a cost of

---

[3] If $r_s = r$, then the results of this paper will not hold, and the lack of Pareto optimality in the socialist state will be limited to the fact that the arbitrarily determined rate of savings does not coincide with the optimum rate of savings in a time-preference sense Lange acknowledges this problem but denies its importance—surely a patent on that line of debate would be worth a fortune It is of course ridiculous for $r_s > r$ in a socialist state

[4] The following argument shows that the failure of Pareto optimality follows from the divergence of r and r_s, and it makes no difference to the failure which particular subset of prices properly reflects social marginal cost, if any The socialist-firm prices are assumed optimal only to abstract from the question of optimal intertemporal choice by assuming that the government-determined rate of time preference is socially correct

[5] It is not clear if borrowing is allowed The case of no borrowing will be considered below

$$C_1(X_2) = p_1X_1 + (r_s + \delta)p_3X_3. \qquad (7)$$

He will want to produce any level of X_2 in the cheapest way, so form the Lagrangian expression

$$V = p_1X_1 + (r_s + \delta)p_3X_3 - \lambda[X_2 - f(X_1, X_3)] \qquad (8)$$

The first-order conditions for a minimum are

$$\frac{\partial V}{\partial X_1} = p_1 - \lambda \frac{\partial f}{\partial X_1} = 0; \qquad (9)$$

$$\frac{\partial V}{\partial X_3} = (r_s + \delta)p_3 - \lambda \frac{\partial f}{\partial X_3} = 0, \qquad (10)$$

$$\frac{\partial V}{\partial \lambda} = X_2 - f(X_1, X_3) = 0. \qquad (11)$$

Since λ is the shadow price of X_2, substitute $p_2^s = \lambda$ in equation (10) and rearrange terms to get the consumer relationship [6]

$$p_2^s = \frac{r_s + \delta}{\partial f/\partial X_3} p_3 \qquad (12)$$

Since r_s is less than r, either (1) the shadow price p_2^s must be less than the firm price p_2, or (2) the household marginal product of X_3 must be less than the firm marginal product, or (3) both must occur [7] Since the social marginal rate of transformation between X_2 and, say, good X_i, produced by the socialist firms,[8] is given by p_2/p_i, there will be failure of Pareto optimality In case 1, this is because the consumer is consuming according to inappropriate price ratios so that there is inefficiency in consumption (a shift of resources from the household production of X_2 to the firm production of X_i would result in increased utility) In case 2, there is inefficiency in production since a shift of production of X_2 from the household to the firms would result in increased output In case 3, there is both inefficiency in consumption and in production Summarizing, in cases 1 and 3, consumers shift savings from the savings bank and consumption from the firm-produced output to consumption of service flows from consumers' durable goods, which are made relatively attractive by the divergence of r_s from r.[9] In cases 2 and 3, there will be inefficient production in the household so that the marginal return to investment is reduced to or toward r_s. Not surprisingly, the closure of the market to private producers leads to increased nonmarket production and the resulting inefficiencies.

---

[6] A similar discussion could be made about inefficiencies in the use of X_1, but as it need not be an argument of the production function, that argument is left aside

[7] These cases correspond to the output effect, substitution effect, and both effects in the theory of production I exclude the possibility that X_3 is an inferior factor

[8] For some products, the optimum scale of output will be sufficiently large that household production would not be worthwhile despite the lower cost of capital to the household than the firm. This is particularly likely to be the case where capital investment is small relative to total costs.

As noted above, it is not clear whether borrowing is to be allowed so consider also the case in which it is prohibited.[10] For those individuals who would still have savings in the state bank, the analysis is unaltered. For those who would want to borrow at r_s but not at $r_l < r$, the analysis is altered to the extent that the internal rate of discount is increased to r_l so that savings in the form of consumers' durable goods yields a return at the rate r_l, but otherwise is unaffected. Only for those individuals who would borrow even at a rate of interest, r, does Pareto optimality hold, since they purchase only the rental services and produce no X_2 themselves.[11]

If the last class contains all citizens—so that at the market rate of interest no one wishes to save even in the form of durables—then Pareto optimality with respect to current consumption and production is achieved. Comparison with the real world makes me, at least, doubt that, in that case, the intertemporal distortions would be of the second order of magnitude.

Another way around the problem might be to sell only rental services. These must not be leased over more than one period of course. But then the comparison with the capitalist state is rather altered, since there is reason to believe that—at least for some items such as automobiles or clothes—rental of services is considerably less efficient than consumer ownership.[12]

Basically we have been talking about distortions due to allowing consumers to save by investing in market goods which alters either their production or demand conditions or both. There is no basic difference in this respect between an investment in human capital and an investment in a durable good. So even were the socialist state to get around the problem of consumers' durable goods, there is still the problem of human capital. Thus the state would have to determine each individual's education and occupation—and this is not only a contradiction with Lange's assumption of "freedom of choice of occupation" (1938, p 83), but returns to the forefront the requirements of huge (impossible?) amounts of information on the part of the central planning board If socialism is to be efficient even in the single period, then control over all produced means of production, including human capital, is required, and this is backbreaking both in terms of information requirements and freedom

---

[9] It is an interesting question, which need not be considered in detail here, whether investment for other purposes is to be affected by a shift of production of services of consumers' durable goods from the firms to the household sector. It would appear that some sort of "production illusion" is present if it is allowed to increase total social investment, but this problem has apparently not been considered by socialist writers.

[10] Borrowing at a different interest rate than r_s is a straightforward extension of the analysis.

[11] Only those who would neither borrow nor save at exactly r might produce some X_2 themselves.

[12] Alchian and Demsetz (1972), discuss some aspects of this problem.

References

Alchian, Armen A , and Demsetz, Harold "Production Information Costs and Economic Organization " *A.E R.* 62 (December 1972) 777–95

Bator, Francis M "The Anatomy of Market Failure " *Q J E* 72 (August 1958) 351–79

Lange, Oskar. "On the Economic Theory of Socialism " In *On the Economic Theory of Socialism,* edited by Benjamin E Lippincott Minneapolis Univ Minnesota Press, 1938

Lerner, Abba P *The Economics of Control* New York Macmillan, 1946

Little, I M. D *A Critique of Welfare Economics* 2d ed Oxford Oxford Univ. Press, 1957

Roberts, Paul Craig "Oskar Lange's Theory of Socialist Planning " *J.P.E* 79 (May/June 1971) 362–77

A Test of the Permanent-Income Hypothesis: Comment

James M Holmes

State University of New York at Buffalo

In a recent article in this *Journal*, Prem Laumas (1969) proposes a new test, which has been repeated elsewhere by Laumas and Mohabbat (1972), of Friedman's permanent-income hypothesis (PIH), using Canadian national income data He concluded that this hypothesis should be rejected In the present paper, we shall demonstrate that Laumas's test is invalid on both a theoretical and empirical basis In the Laumas model, it is possible to test seven assumptions which constitute Friedman's hypothesis In fact, five of Friedman's seven assumptions are violated in Laumas's tests of the remaining two Thus, even if Laumas's theoretical analysis were saisfactory, it could not be used to reject Friedman's theory Moreover, Laumas's empirical evidence suffers from one major and one minor defect Finally, we present a rigorous direct test of Friedman's hypothesis which violates none of the PIH assumptions using exactly the same Canadian national income data as used by Laumas, except that our data will be deflated This direct test leads to a rejection of the PIH when consumption is defined to include consumer-durable expenditures However, when consumption is defined to exclude such expenditures, this part of Friedman's theory is not easily rejected on the basis of the Canadian evidence This result is in sharp contrast to Laumas's claim and to that obtained from the United States evidence previously reported by Holmes (1970)

I. Friedman's Theory and Laumas's Test

Friedman (1957, chap 3) formulated the following model, where C_p = permanent consumption, C_T = transitory consumption, C = measured consumption, Y_p = permanent income, Y_T = transitory income, and Y = measured income All variables are contemporaneous unless indicated otherwise by a bracketed t The term ρ_{XY} is the theoretical

---

I wish to acknowledge the aid of my research assistant Emmanuel Tzanakis, Bryon Spencer of McMaster University for his help with the data, and Donald Ebbeler, David Smyth, and Nagesh Revankar for their comments.

correlation between X and Y

$$C_P = kY_P, \tag{1}$$

$$Y = Y_P + Y_T, \tag{2}$$

$$C = C_P + C_T, \tag{3}$$

$$\rho_{Y_T Y_P} = 0, \tag{4}$$

$$\rho_{C_T C_P} = 0, \tag{5}$$

$$\rho_{Y_T C_T} = 0 \tag{6}$$

From equation (1), we infer that transitory income has no effect upon permanent consumption From equations (1) and (4), we deduce equation (7), and from equations (1) and (5), we deduce that permanent income has no effect on transitory consumption. Thus,

$$\rho_{C_P Y_T} = 0, \tag{7}$$

$$\rho_{C_T Y_P} = 0 \tag{8}$$

Although Friedman did not specify the distribution of the error term C_T in the regression of C on Y_P, his statistical tests implicitly assumed that C_T was distributed normally with expectation zero and constant variance

$$C_T \sim \Phi(0, \sigma_1^2) \tag{9}$$

Friedman ran the regression of C on Y_P which is derived from equations (1) and (3)

$$C(t) = kY_P(t) + C_T(t) \tag{10}$$

The estimates of the marginal propensity to consume k, granted the assumptions above, will be best linear and unbiased if $Y_P(t)$ is defined Friedman defined $Y_P(t)$ (1957, pp 142–46) as

$$Y_P(t_0) = \beta \int_{-\infty}^{t_0} e^{(\beta-\alpha)(t-t^0)}\, Y(t)\, dt, \tag{11}$$

where β is the coefficient determining expectations adjustments and α is the secular rate of growth of permanent income [1] Friedman assumed $\beta = 4$ and $\alpha = 02$

The model Laumas uses to test Friedman's theory consists of basically two regression equations

$$C_P^L = \alpha_0 + \alpha_1 Y_P + \varepsilon, \tag{12}$$

$$C_T^L = B_0 + B_1 Y_T + V, \tag{13}$$

---

[1] I have demonstrated that this assumption contradicts the previous assumptions. The theoretical defects of the PIH may, in fact, be more limiting than the empirical, since they may apply to any use of the permanent income concept (see Holmes 1971). For exceptions, see Ebbeler (1971).

where C_P^L and C_T^L are permanent and transitory consumption as defined by Laumas The ε and V are stochastic-error terms, which we may infer on the basis of Laumas's statistical tests (1969, p 859), are each normally distributed with expectation zero and constant variance The measure of permanent consumption is given by a scheme similar to permanent income

$$C_P^L(t_0) = \lambda \int_{-\infty}^{t_0} e^{\lambda(t-t^0)} C(t)\, dt \tag{14}$$

Laumas assumed $\lambda = 6$ for consumption defined both to include and exclude expenditures on consumer durables [2] For these two differing definitions of consumption and thus permanent consumption, he assumed, respectively, that $\beta = 45$ and $\alpha = 0$ and $\beta = 35$, $\alpha = 0$ in equation (11) defining permanent income Thus, equations (12) and (13) apply for two definitions of consumption and permanent income

Laumas claims that Friedman's theory implies that $E(\alpha_1) > 0$ and $E(B_1) = 0$ in equations (12) and (13) (i e , in terms of correlation coefficients, the assumptions in eqq [1] and [6]) He tested these two statements and found the latter unacceptable on the basis of the evidence from the Canadian economy

Friedman's theory does not imply that $E(B_1) = 0$ Equation (12) is not equivalent to equation (1), that is, $C_P^L \neq C_P$ Permanent consumption according to Friedman is estimated to be exactly proportional to permanent income Laumas assumes in equation (14) that C_P^L is estimated as a moving average of past consumption and that this estimate contains an error, that is, $E(\varepsilon) \neq 0$ Thus, in equation (13), $C_T^L \neq C_T$ According to Friedman, C_T is also not estimable independent of income, a result contrary to Laumas

It is possible to prove, using the model in Holmes (1973), that $C_P^L \neq C_P$, $C_L^T \neq C_T$, $C_P^L > Y_P$, whereas Friedman's theory always does well in "explaining" consumption, that is, $C_P(t) = C(t)$ and $C_T = 0$

The empirical evidence presented by Laumas contains one major error the weights used to construct permanent income sum to less than unity [3] Whereas, if there is secular growth in income, that is, $\alpha > 0$, these weights should sum to more than unity If one regresses current consumption including durables in either real or nominal terms on either permanent

[2] Laumas (1969, p 859 n) states "We finally chose disposable income series with $\beta = 35$ and $\beta = 45$ by using consumption series excluding durable goods, and consumption series including durable goods, respectively, by the goodness of fit criteria Permanent consumption was also computed similarly The β coefficients turned out to be 6 for both consumption series (including and excluding durable goods) "

[3] Laumas states (ibid): "With $B = 35$, fourteen years (1930–43) were used up in computing permanent income With $B = .35$, the weights used were (4 decimal places) 2953, .2081, .1466, 1033, .0728, .0513, 0362, 0255, 0180, 0127, 0089, .0063, 0044, 0031, .0022, .0015, and .0011 The sum of these weights equals unity " The sum of the first 15 weights is .9973 − .0026 = .9947 These 15 weights should be normalized, and α should be > 0.

TABLE 1
SOME ESTIMATABLE STATISTICS IN THE LAUMAS VERSION OF PIH

$N = 23$	Including Durables $B = 45$	Excluding Durables $B = 35$
$\hat{E}(C_T^L)$	1,463 21	1,242 17
Standard error	(559 49)	(472 72)
$\hat{\rho}_{C_T^L C_P^L}$	8424	8396
$\hat{\rho}_{C_T^L Y_P}$	8458	8430
$\hat{\rho}_{C_P^L Y_T}$	8519	8763
$\hat{\rho}_{Y_P Y_T}$	8593	8815

NOTE —The data used were Canadian National Accounts series 30014, 30033, 30300, 30340 30972, and 31058, kindly supplied to me by Bryon Spencer of McMaster University, and with the exception of the price index correspond to Laumas's data The weights used in constructing the moving averages are corrected to sum to unity, however

income using $\beta = 35$ or $\beta = 45$, after the weights of such series are normalized, the marginal propensity to consume is always greater than one Moreover, for several years in the 1960s, consumption was greater than permanent income using Laumas's weights

The direct empirical evidence concerning the validity of the Laumas tests supports our theoretical criticism of these tests In Friedman's regression equation (10), the estimates of $E(C_T)$, $\rho_{C_T C_P}$, and $\rho_{C_T Y_P}$ are constrained to be zero In the Laumas model, equations (12) and (13), these statistics can be estimated Moreover, we can estimate $\rho_{C_P^L Y_T}$ and $\rho_{C_T Y_P}$ in the Laumas model Table 1 presents the estimates of these statistics using the same nominal Canadian national income data and time period (1944–66) as Laumas uses

One must conclude on the basis of this evidence that the assumption that $E(C_T^L) = 0$ is not acceptable and that the correlation coefficients are significantly different from zero at the 01 level of significance Clearly, if a statistical test of a hypothesis violates any of the assumptions under which that hypothesis is supposed to hold, the test is invalid In the case at hand, every assumption of the PIH which is not tested (i e , $\rho_{C_P Y_P} > 0$ and $\rho_{C_T Y_T} = 0$) is violated The fact that $\hat{\rho}_{C_T^L Y_T} > 0$ cannot be considered as evidence upon which to reject (or accept) the PIH under these circumstances

There is further theoretical difficulty in this test The estimated correlation between permanent and transitory income is positive As I demonstrated elsewhere (1971), under certain circumstances, the theoretical correlation between these two variables has a positive expectation when the permanent series is defined by an equation such as (11) or (14). Thus, we can expect that, under such circumstances, $\rho_{C_P^L C_T^L} > 0$ and $\rho_{Y_P Y_T} > 0$ According to Friedman's theory, $\rho_{C_P Y_P} > 0$ Thus, if C_T is a stochastic term as Friedman assumed, then equations (12) and (13) each contain part of the effect of C_P in the dependent variable. This may explain why

TABLE 2
ALTERNATIVE REGRESSIONS OF CONSUMPTION ON BOTH TRANSITORY AND PERMANENT INCOME

	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	D-W	R^2
A Consumption including durables·					
1 $Y_P^*(\beta = 35), \hat{\rho} = 0$	−3,036	261	1.130	658	.993
	(446 3)	(244)	(0 034)		
2 $Y_P^*(\beta = 35), \hat{\rho} = 6030$	−2,380	137	1 116		981
	(838 9)	(166)	(0 044)		
3 $Y_P^*(\beta = 45), \hat{\rho} = 0$	−3,077	215	1 116	749	993
	(464 6)	(283)	(0 032)		
4 $Y_P^*(\beta = 45), \hat{\rho} = 6045$	−2,326	.077	1 097		981
	(832 5)	(176)	(0 041)		
5 $Y_P^*(\beta = 4, \alpha = 035), \hat{\rho} = 0$	−3,141	242	1 056	.8672	9937
	(444)	(235)	(0 019)		
6 $Y_P^*(\beta = 4, \alpha = 035), \hat{\rho} = .4965$	−2,306	345	1 024		9856
	(714)	(160)	(0 029)		
B Consumption excluding durables					
7 $Y_P^*(\beta = 35), \hat{\rho} = 0$	−806 8	304	0 889	804	994
	(318 7)	(174)	(0 024)		
8 $Y_P^*(\beta = 35), \hat{\rho} = 5341$	−484 6	134	0 903		985
	(542 7)	(129)	(0 030)		
9 $Y_P^*(\beta = 45), \hat{\rho} = 0$	−852 5	223	0 893	749	994
	(321 5)	(196)	(0 023)		
10 $Y_P^*(\beta = 45), \hat{\rho} = 5528$	−469 2	065	0 890		988
	(540 8)	(132)	(0 028)		
11 $Y_P^*(\beta = 4, \alpha = 035), \hat{\rho} = 0$	−880	288	(0 848)	9031	99504
	(316)	(168)	(0 0136)		
12 $Y_P^*(\beta = 4, = 035), = 4829$	−383	272	0 828		98859
	(513)	(119)	(0 0210)		

NOTE —In the first column the value of β in parentheses refers to the weights used to construct the Y_P^* series In the second column the value of $\hat{\rho}$ is given which was utilized in conjunction with eq (16) to estimate the regression equation Figures in parentheses, in columns 3–5, represent standard errors.

$\hat{\alpha} > 0$ It also implies that, if $\hat{B}_1 > 0$, we cannot distinguish between this being caused by $\rho_{Y_TC_P} > 0$ or $\rho_{Y_TC_T} > 0$

II. A Correct Test of Friedman's Theory

The PIH summarized in equations (1) through (11) implies in equation (15) that $\gamma_0 = 0 = \gamma_1$ and $1 > \gamma_2 > 0$ We utilize real instead of nominal values of these variables following Friedman We star such variables We can test if $\rho_{Y_P^*Y_T^*} = 0$

$$C^* = \gamma_0 + \gamma_1 Y_T^* + \gamma_2 Y_P^* + C_T^*. \tag{15}$$

The estimates of equation (15), assuming C_T^* is distributed as in equation (9) and alternatively assuming that

$$C_T^*(t) - \rho C_T^*(t-1) \sim \Phi(0, \sigma_2^2), \tag{16}$$

are presented in table 2 We estimated equation (15) using the two measures of consumption and the two measures of permanent income used by Laumas after the weights were normalized. In addition, we constructed a permanent income series using the weights derived for $B = .4$ and $\alpha = .035$, which is an estimate of the postwar growth in Canadian income In this latter case the sum of the weights exceeded unity; they equated approximately 1 096 [4]

We note that Friedman's PIH must always be rejected when the dependent variable includes durables expenditure in the first six equations First, because $\hat{\gamma} \neq 0$ for any specification of Y_P^* or for either specification of the disturbance term, C_T^* (the lowest t-value is $-2\ 794$ in line 4) More critical to Friedman's theory is the fact that $\hat{\gamma}_2$ is significantly greater than one at .05 level in the first five equations Friedman's theory cannot be rejected on the basis that the marginal propensity to consume out of transitory income is different from zero ($\hat{\gamma}_1$) in these first five equations However, in equation (6), the hypothesis that transitory income has no effect upon permanent consumption must be rejected Equation (6) is, in my opinion, the one most relevant to the PIH, as it seems silly to assume the expected secular rate of growth of income in postwar Canada was zero

When real consumption excluding durables expenditures is the dependent variable, the PIH is rejected when it is assumed, following Friedman, that $C_T^* \sim \Phi(0, \sigma_1^2)$ since $\hat{\gamma}_0$ is significantly less than zero at the 05 level. If one is willing to drop this portion of Friedman's theory, then the estimates of the marginal propensities to consume out of permanent and transitory income are in good agreement with the PIH However, the D-W statistic indicates considerable serial correlation in the residuals of these three equations (lines 7, 9, and 11) When the error term C_T^* is assumed to be generated from a first-order Markov process, Friedman's PIH cannot be rejected if we use Laumas's weights, that is, if one assumes the expected secular rate of growth of income is zero However, if we do assume that the secular rate of growth of the real income of Canada is 3 5 percent, our results (eq [12]) are similar to those obtained for the United States economy. transitory income has a significant effect on permanent consumption

Last, we can test if $\rho_{Y_T^* Y_P^*} = 0$. For $\beta = .35$ $\rho_{Y_P^* Y_T^*} = 0\ 7856$ and $= 7595$ for $\beta = .45$, both of which imply that this assumption is not acceptable at the usual level of significance. It is a result consistent with our theoretical expectations Holmes (1971)

In conclusion, we have demonstrated that Laumas's new measure of permanent consumption does not correspond to Friedman's Moreover, we showed that Laumas's test of the PIH violated all of the assumptions underlying that hypothesis which were not being tested. In the second section of this paper, we presented a test of Friedman's PIH consistent with his assumptions and found that, if consumption is defined to exclude durable expenditures, the experience of the Canadian postwar economy could be considered in good agreement with this hypothesis if we follow Laumas and assume $\alpha = 0$ However, if we assume that $\alpha = .035$, in our opinion a more reasonable assumption, we find that transitory income is a significant explanatory variable. We can conclude on the basis of this evidence that there may be a critical difference empirically between the role played by transitory income and the expected rate of secular growth of income in the United States and Canadian economies

---

[4] The fact that as $\alpha \gtreqless 0$ the sum of the weights should exceed, equal, or be less than unity, as I demonstrate in Holmes [1973], is not widely recognized by previous researchers. In my opinion, there appears to be both on theoretical and empirical grounds sound basis for considering the role of α, the expected secular rate of growth of income more critical to the PIH than that of β, the coefficient of expectations adaptation.

References

Ebbeler, D. H. "On the Measurement of the Permanent Component of a Series " Claremont Economic Papers no. 6 (July 1971), revised in *J American Statis Assoc*, in press.

Friedman, H *A Theory of the Consumption Function.* Princeton, N J. Princeton Univ Press, 1957.

Holmes, J. M. "A Direct Test of Friedman's Permanent Income Theory " *J. American Statis Assoc.* 65 (September 1970) 1159–62

——— "A Condition for Independence of Permanent and Transitory Components of a Series." *J American Statis. Assoc* 66 (March 1971) 13–15

——— "The Role of Secular Growth in the Permanent Income Hypothesis " Discussion Paper no 274, State Univ New York at Buffalo, 1973

Laumas, Prem S. "A Test of the Permanent Income Hypothesis." *J P.E* 77, no 5 (September 1969) 857–61

Laumas, Prem S., and Mohabbat, Khan A "The Permanent Income Hypothesis Evidence from Time-Series Data " *A E R* 62 (September 1972) 730–35.

A Test of the Permanent-Income Hypothesis: Reply

Prem S. Laumas

Northern Illinois University

Holmes raises three main issues with respect to my (1969) test of the permanent-income hypothesis (PIH): (*a*) that my definition of permanent consumption is not equivalent to that of Friedman, (*b*) that the weights used to derive permanent income and permanent consumption do not sum to unity and hence the values of permanent income, permanent consumption and transitory income, and transitory consumption used in my test are not the "correct" values, (*c*) that I ignored the average-trend rate of growth in calculating permanent income.

Before I consider these issues, let me state that the conclusions arrived at by Holmes are not inconsistent with mine. He states: "This direct test leads to a rejection of the PIH when consumption is defined to include consumer durable expenditure. However, when consumption is defined to exclude such expenditures, this part of Friedman's theory is not easily rejected on the basis of Canadian evidence" (1974).

Holmes himself points out that his equations (6) and (12) in table 2 are the really relevant equations. The results of these equations lead to a mild rejection of the PIH for Canada. Thus, according to equation (6) of table 2, when consumption is defined to include durable goods, the marginal propensity to consume out of transitory income is equal to .345, and the marginal propensity to consume out of permanent income is equal to 1.024. When consumption excludes durable goods (according to eq. [12] of table 2), the marginal propensity to consume out of transitory income is equal to .272 and out of permanent income is equal to .828. Since consumption, according to the PIH, is defined to include the value of all nondurable goods plus the rental value of durable goods, one might expect that the true marginal propensities to consume transitory and permanent incomes would lie within these limits. It is likely that the true marginal propensities to consume out of transitory and permanent incomes would be approximately .30 and .90, respectively. These results are not very different from those that I reported in my paper and are very similar to those reported in tables 2 and 3 below.

Let us now turn to the three issues raised by Holmes.

The author is thankful to Gurcharan Laumas and Eldon Reiling, for helpful comments on an earlier draft of this paper.

Definition of Permanent Consumption

Holmes argues that my definition of permanent consumption is not the same as Friedman's I do not fully agree with this I made several series of permanent consumption, according to different values of β, and chose that particular series of permanent consumption (C_p) which showed the highest correlation with the permanent income (Y_p) series which had already been selected I justify this procedure on the ground that I provisionally assume that Friedman is right in that C_p and Y_p are most highly correlated and then try to test if the proportionality hypothesis and its most important assumption, namely, the supposed marginal propensity to consume out of transitory income being zero, is valid or not A series of C_p generated in this way is unacceptable to Holmes Rather, he suggests that according to PIH, C_p is by definition proportional to Y_p I cannot agree with Holmes on this point Obviously, one can calculate $\hat{C}_p$ as $\hat{k}Y_p$ (as reported below) If the parameter k is truly constant, then C_p is perfectly correlated with Y_p, not just highly correlated, as I assumed This method of calculating $\hat{C}_p$ is acceptable to Holmes But is this a correct test of the PIH? My answer is no How can one claim to test a hypothesis when the testing procedure is predesigned to validate the hypothesis?

Friedman's PIH comes out with flying colors if one were to assume that permanent consumption is proportional to permanent income, that is, $\hat{C}_p = \hat{k}Y_p$ Permanent consumption was found by taking the estimated marginal propensity to consume and multiplying it into the constructed permanent-income series. Transitory income and transitory consumption were found by deducting permanent income and permanent consumption from their respective measured values Estimates were made for the following equations

$$\hat{C}_p^* = \alpha_0 + \alpha_1 Y_p + \mu_1, \tag{1}$$

$$\hat{C}_p = \beta_0 + \beta_1 Y_p + \mu_2, \tag{2}$$

$$\hat{C}_t^* = \gamma_0 + \gamma_1 Y_p + \mu_3, \tag{3}$$

$$\hat{C}_t = \delta_0 + \delta_1 Y_p + \mu_4, \tag{4}$$

where $\hat{C}_p^*$ is aggregate permanent consumption, including durable goods, and $\hat{C}_p$ is permanent consumption excluding durable goods, $\hat{C}_t^*$ and $\hat{C}_t$ are the corresponding transitory consumptions, Y_p and Y_t represent permanent and transitory incomes, respectively, and μ_1, μ_2, μ_3, and μ_4 are the error terms of the regression equations In table 1, we present the results on this basis, weights are normalized to sum to unity, and, following Holmes, α was assumed to be 3 5 percent

The values of the MPC_t are not just exceedingly small, but also insignificant at the 5 percent level The correlation between the transitory

TABLE 1

MARGINAL PROPENSITIES TO CONSUME PERMANENT AND TRANSITORY INCOMES FOR CANADA (1944–66)

	MPC_p	MPC_t	$\gamma_{y_t c_t}$
(1) $C^* = f(Y)$	.99132	.1367 (.1868)	.244
(2) $C = f(Y)$	88801	0926 (.0716)	.271

NOTE —Figures in parentheses indicate standard errors of the coefficients. Intercepts of all equations were insignificant C^* = consumption including durable goods, C = consumption excluding durable goods, MPC = marginal propensity to consume Subscripts p and t represent permanent and transitory variables, y = disposable income, γ = coefficient of correlation. Estimates of MPC_p are significant at the 5% level

components of income and consumption is also very small Clearly, the most important assumption of Friedman's PIH is valid if we grant that permanent consumption is proportional to permanent income.

The Weighting Pattern

Holmes makes the point that there is a serious error in my calculations insofar as the weights used to derive permanent income do not sum to unity. Thus, when β = 35, my weights add up to 9973 It appears to be a very weak criticism of my paper. It is no doubt true that weights should, ideally, sum to unity Ideally, if one were to go back several more years beyond what I did, no doubt the sum of the weights would add to unity But, would that have been a worthwhile exercise? In part, Friedman answers this question with respect to his own calculations "The retention of as many as 17 terms is doubtless an excess of precision. It is dubious that the results would be appreciably affected by retaining, say, only 9 terms and adjusting the weights for them to sum to unity "[1]

Holmes is right in stating that if the weights do not sum to unity there is a bias introduced in the calculation of permanent and transitory incomes, and, therefore, in the calculations of marginal propensity to consume out of permanent and transitory incomes It may be added, however, that if one normalizes the weights, there is another type of bias introduced into the calculation of permanent and transitory income due to the normalization factor Values of permanent and transitory incomes may go slightly up or slightly down depending on how the weights were normalized.

However, I recalculated my estimates for permanent and transitory variables with normalized weights and found that the new results are very similar to the ones reported in my original paper Thus, for example, when the weights are normalized, the MPC_p including durable goods was 970, as compared with .968 when the weights were not normalized. Again, the MPC_p excluding durable goods was 867 when weights were normalized and .871 when the weights were not normalized. These results, in part, support Friedman's intuitive hunch reported above.

---

[1] Friedman (1957, p. 146, n. 26).

Parenthetically, we may note that Holmes's view of the role of weighting scheme in the PIH makes one uncomfortable. Throughout his paper, he states that Friedman assumed $\beta = .4$ and $\alpha = .02$ and that I assumed certain β values in my calculations. This is not true. Neither Friedman nor I made such assumptions, but found these weights from the data sets In his own equations (6) and (12) of table 2, which he considers as the most satisfactory equations, Holmes states that $\beta = 4$ He tries to justify this "because this was Friedman's original assumption or that it is the average of two alternative assumptions of β assumed by Laumas " (Holmes 1974). I believe that this is arbitrary and an unscientific procedure The value of β ultimately determines the length of the "horizon" which is an integral part of the PIH The "horizon" in turn reflects the state of economic development of a country, its history, economic structure, and how people form their expectations In the absence of any comparison between the United States and Canada along these lines, I feel that Holmes is incorrect in following Friedman and assuming $\beta = 4$ for Canada Again, it appears strange that Holmes should consider an average of the two values of β that I reported in my original paper for calculating permanent income relevant to the consumption of all goods, and consumption excluding durable goods as somehow a more appropriate figure for estimating consumption functions including durable goods and excluding durable goods in testing the validity of the permanent-income hypothesis for Canada There is no justification for this in his paper

A hypothesis as elegant and as powerful as the PIH appears rarely in the literature It is unfair to the author of such a theory to accept without reservations the central tenets of such a theory The exact value of β, and thus the length of the horizon, is still a controversial issue in the literature[2] It influences the extent to which the permanent value of a variable is likely to differ from its measured value, and thus the whole question of the validity of the distinction between permanent income and transitory income.

Use of the Trend Factor

Without the trend factor (α), an equation like equation (11) in Holmes's paper (1974), when applied to a steadily growing series, yields estimated values of a permanent variable systematically below their measured values. If one were to suppose that the PIH in its strict version is valid, it makes no difference if proper allowance is made for the trend factor

---

[2] See, for example, Tobin and Swan (1969), Wright (1969), Holbrook (1967), and Landsberger (1971)

TABLE 2
MARGINAL PROPENSITIES TO CONSUME PERMANENT AND TRANSITORY INCOME IN CANADA (1944–66)

	MPC_p	$\gamma_{y_pc_p}$	MPC_t	$\gamma_{y_tc_t}$
(1) $C^* = f(Y)$	977 (.008)	99	.306 (043)	83
(2) $C = f(Y)$	843 (165)	99	367 (047)	86

NOTE —For meaning of symbols, please see note at the end of table 1 Intercepts of MPC_p with and without durable goods were insignificant at the 5% level, while those of the MPC_t equations were significant at the 5% level

However, following Holmes, by using a value of $\alpha = .035$ for both Y_p and C_p and by normalizing weights, I present the new results in table 2 for equations (1)–(4) above Permanent consumption was computed as in my 1969 paper and as explained above, and not as $\hat{C}_p = \hat{k}Y_p$

Results of table 2 indicate that the true MPC_t would be approximately one-third of the true MPC_p It may be noted that this modifies my results reported earlier (1969) There I had found MPC_t to be approximately one-half of MPC_p Thus, I wish to point out that the PIH is more valid for Canada than I thought on the basis of my previous calculations.

Holmes argues that this conclusion needs further modification due to the existence of high correlation between Y_t and C_p The essence of his point is that since C_p^L is an imperfect measure of the true C_p and since my estimation procedure leads to a certain systematic element of the true C_p being transferred to C_t^L, we do not know if my estimated MPC_t is due to $\gamma_{y_tC_t}$ or $\gamma_{y_tC_p}$ He suggests that at least part of MPC_t may be due to high $\gamma_{y_tC_p}$ In my calculations relating to table 2, I find $\gamma_{y_tC_p^*} = 768$ and $\gamma_{y_tC_p} = 839$ If one were to grant Holmes's point, one would have to modify the conclusion that MPC_t in Canada is approximately one-third of MPC_p However, in the absence of a valid method of calculating the degree of upward bias in MPC_t (due largely to our ignorance of the true C_p), one does not know if our MPC_t and MPC_p values differ significantly from their true counterparts[3]

In addition to the above tests, following Holmes, I made a direct test of the PIH in which measured consumption is the dependent variable and the transitory and permanent components of income the explanatory

---

[3] Athough Holmes's argument appears sound in theory, from an empirical point of view it seems inconclusive. Thus, with respect to our calculations in table 1, when $\hat{C}_p$ is defined as $\hat{k}Y_p$ (this makes $\hat{C}_p$ as close to being its true value as possible), we find that $\gamma_{y_tC_p^*} = 774$ and $\gamma_{y_tC_p} = .841$ (These values are somewhat higher than when C_p^L is used as a measure of permanent consumption) In this case, therefore, the existence of a high $\gamma_{y_tC_p}$ appears quite consistent with the near validity of a strict version of the PIH for Canada In further view of this result, one does not know how far one must alter, if at all, the conclusions arrived at from figures in table 2.

TABLE 3
MARGINAL PROPENSITIES TO CONSUME OUT OF PERMANENT AND TRANSITORY INCOMES FOR CANADA (1944–66)

	R
$C^* = 27\ 42 + 970\ Y_p + 324\ Y_t$.	.99
[.365] [66 509] [1 866]	
$C = 330\ 99 + 859\ Y_p + 316\ Y_t$	99
[5 84] [64 881] [2.598]	

NOTE —Figures in brackets are the t-values The C^* is measured consumption including durable goods, and C is measured consumption excluding durable goods

variables Estimates were made for the following equation

$$C^* = \lambda_0 + \lambda_1 Y_p + \lambda_2 Y_t, \tag{5}$$

$$C = \partial_0 + \partial_1 Y_p + \partial_2 Y_t, \tag{6}$$

where C^* is aggregate measured consumption including durable goods and C is measured consumption excluding durable goods The series of Y_p used here is the same as in the previous tests for table 2 The results are presented in table 3

The results of table 3 appear consistent with the results from table 2 The marginal propensity to consume out of transitory income appears to be approximately one-third of the MPC_p

One concluding observation appears in order Although I have carefully followed Friedman's method for calculating the permanent value of a variable, I do not fully agree that his methodology corresponds most accurately to his theoretical construct of the concept of permanent income True, there is considerable valid psychological reasoning behind this method,[4] yet it appears to me that the assumption that the biggest response occurs immediately at the beginning of the adjustment period is rather strong

To summarize, then, the following points emerge from the above discussion (*a*) Holmes's results seem consistent with mine when I make due allowance for trend and normalize weights, (*b*) Friedman's PIH does not need as much modification in its application to Canada as I had suggested in my original paper; (*c*) Holmes's criticism of my concept of C_p is based on what I consider an erroneous view of hypothesis testing, (*d*) Holmes's paper does not show adequate appreciation of the importance of "horizon" in the PIH, (*e*) both Holmes's test and my test are based on a particular method of calculating the permanent value of a variable which may not be a perfect method, and it is quite possible that a better method may lead to more accurate results.

[4] The psychological postulates as related to Jost's law of memory, underlying Friedman's theory, are briefly discussed in Laumas and Laumas (1972).

◆

References

Friedman, M *A Theory of Consumption Function.* Princeton, N.J.: Princeton Univ. Press, 1957.

Holbrook, R S. "The Three-Year Horizon An Analysis of the Evidence." *J.P.E.* 75, no 5 (October 1967)· 750–54

Holmes, J. "A Test of the Permanent-Income Hypothesis. Comment " *J.P.E* 82, no 1 (January–February 1974). 185–91

Landsberger, M "Consumer Discount Rate and the Horizon New Evidence." *J.P E.* 79 (November–December 1971)· 1346–59

Laumas, P S "A Test of the Permanent Income Hypothesis." *J.P E.* 77, no 5 (September–October 1969) 857–61.

Laumas, P , and Laumas, G S "On How to Calculate Permanent Income " *Rev Income and Wealth* Ser. 18, no 4 (December 1972). 435–38

Tobin, J , and Swan, C "Money and Permanent Income. Some Empirical Tests " *American Econ Assoc. Papers and Proc.* (May 1969), pp. 285–95.

Wright, C "Estimating Permanent Income A Note." *J.P E.* 77, no 5 (September–October 1969) 845–50

"Preferences, Separability, and the Patinkin Model": A Comment

Eitan Berglas and Assaf Razin

Tel-Aviv University

I

It has been shown that the introduction of real balances into the utility function, which is the main feature of the Patinkin model, generally destroys the Slutsky results. More specifically, the derivative of the compensated demand of any good with respect to its own price is not necessarily negative.[1]

In his paper, Lloyd shows that the "work by Baumol, Tobin, Patinkin and others implies that the utility function is separable in a specific way" which we shall reproduce later.[2] Lloyd also shows that if this separability assumption holds, the Slutsky results can be reinstated with some modifications. He then proposes an empirical test of the separability assumption.

Pearce (1958, 1964, chap. 3), in his analysis of demand and savings, has made extensive use of the separability assumption to develop the Slutsky results for this case. However, his analysis did not consider any monetary model, in particular, he did not refer to the Patinkin model.

This note shows that it is possible to apply the Pearce analysis to the Patinkin model. Such an application yields a modified version of the Slutsky results. The modified version and its derivation are much simpler than Lloyd's. Furthermore, this derivation can be subjected to a simple empirical test, while the empirical test of the Lloyd formulation is shown to be impractical.

II

Given Lloyd's separability assumption, the utility function can be written[3]

$$u = u[\phi(x_1, \ldots, x_m), p, M], \tag{1}$$

---

[1] For a discussion of this problem and for references to earlier contributions, see Patinkin (1965, pp. 403–10).

[2] Lloyd (1971, p. 642).

[3] We exclude bonds from the analysis, since bond interest rates are constant throughout, thus the relative price of bonds in terms of money is constant, and the two can be regarded as a composite good. The extension of the analysis to the case where interest rates vary is straightforward.

where $x_1, \ldots, x_m$ are amounts of current commodities, p is the price index, and M is the nominal quantity of money. The M and p are introduced separately in the utility function. Obviously, this includes the form M/p, which Lloyd uses, as a special case.

It is useful to write the demand function for current commodities as $x_i = x_i(p_1, \ldots, p_m, E)$, for $i = 1, \ldots, m$, where E is the total expenditure on current commodities. Differentiating toally, we get

$$\frac{dx_i}{dp_j} = \frac{\partial x_i}{\partial p_j} + \frac{\partial x_i}{\partial E}\frac{dE}{dp_j} \tag{2}$$

Partitioning $\partial x_i/\partial p_j$, we get

$$\frac{\partial x_i}{\partial p_j} = S_{ij} - x_j\frac{\partial x_i}{\partial E} \tag{3}$$

Given the budget constraint $E + M = Y + L$, where Y is income and L is the initial quantity of money (both are given to the consumer), and using equations (2) and (3), we get[4]

$$\frac{dx_i}{dp_j} = S_{ij} - \frac{\partial x_i/\partial Y}{1 - \partial M/\partial Y}\left(x_j + \frac{dM}{dp_j}\right) \tag{4}$$

By the separability assumption, the consumer can be regarded as though he is solving a two-stage maximization problem: (*a*) given prices and $L + Y$, he determines M, (*b*) subject to $E = L - M + Y$ and prices, he maximizes ϕ. The terms S_{ij} are the compensated demand derivatives holding ϕ constant. These terms, therefore, have all the known sign and symmetry properties.

III

Equation (14) in Lloyd's paper has the following disadvantages when compared with equation (4) above. (*a*) With Lloyd's approach, it is necessary to know the price index function—for which no satisfactory specification exists,[5] (*b*) equation (14) includes a variable Q which is not observable. (This variable is solved from eq. [14] of a particular pair of commodities.) Lloyd claims that his equation (14) makes the results of traditional demand theory open to empirical falsification. Given the difficulties we have noted, however, such a claim is at best questionable. Equation (4) of this paper, which depends only on observable magnitudes, can be used for this purpose. An even simpler test of the properties of S_{ij} and the separability assumption can be carried out by using equation (3).

We cannot close without noting that what we have done is implied by

---

[4] Note that $dE/dp_j = -dM/dp_j$ and $\partial E/\partial Y = 1 - \partial M/\partial Y$.

[5] For a discussion of these difficulties, see Patinkin (1965, pp. 405–6).

Pearce (1958; 1964, chap. 3). It is true that Pearce does not refer to the Patinkin model However, this note shows that the Patinkin model is a special case of Pearce's model Using Pearce's approach in a monetary model makes it possible to derive a modified version of the traditional demand properties in a way which can be tested empirically.[6]

References

Berglas, E , and Razin, A "A Note on the Balance of Payments and the Terms of Trade in Relation to Financial Controls " *Rev. Econ. Studies* 39 (October 1972): 511–13

Lloyd, C "Preferences, Separability and the Patinkin Model." *J P.E.* 79 (May/June 1971): 642–51.

Patinkin, D. *Money, Interest, and Prices* 2d ed. New York. Harper & Row, 1965.

Pearce, I. F. "Demand Analysis and the Savings Function." *Econ Rec* 34 (April 1958): 52–66

———. *A Contribution to Demand Analysis* London · Oxford Univ. Press, 1964

[6] The usefulness of this approach can be seen from its application to the analysis of the effects of devaluation on the balance of payments in Berglas and Razin (1972)

A Paradoxon of the Market Mechanism

Gerhard O. Orosel

University of Vienna

1. Introduction and Results

The market mechanism can be used to construct a social welfare function in such a way that equilibrium points are preferred to other Pareto-optimal points, which in turn are preferred to non-Pareto-optimal points. K. J. Arrow's condition of positive association (Arrow 1963) is not applicable to this social welfare function because of formal reasons, but similar conditions of the same philosophical content can be formulated. It can be shown by examples that the market mechanism of a barter economy violates these conditions, that is, changes in individual preferences may lead to exactly opposing changes in social preferences.

2. The Market Mechanism as a Social Welfare Function

The market in a barter economy can be looked at as a mechanism which assigns to each given set of utility functions for individual i, $i = 1, 2, \dots, n$, and initial endowments of each individual i with each commodity j, $j = 1, 2, \dots, m$ at least one matrix (a_{ij}), where a_{ij} gives the amount of commodity j individual i gets in equilibrium (cf. Debreu 1959 or Arrow and Hahn 1971). In order to make things easier, there is no production included in this model.[1]

First a social welfare function, induced by the market mechanism, must be constructed. For a social welfare function, we need a complete pre-ordering of the alternatives. Alternatives are now all attainable (feasible or possible) allocations of the commodities, where attainable (feasible or possible) means (1) that for each commodity the sum of the individual quantities of this commodity is constant and equal to the initial endowment of the society with this commodity and (2) that no individual gets less of every commodity in an attainable alternative than in its initial

---

. I am indebted to K. J. Arrow for very valuable discussion and to an unknown referee for most helpful critique.

[1] As Rader (1964) has shown, any model of an economy involving production and exchange can be represented by an equivalent model involving only exchange, if the conditions for an efficient price system are fulfilled.

endowment[2] This follows from the absence of production. Hence, alternatives are matrices (x_{ij}), where x_{ij} is the amount of commodity j individual i gets in alternative x, $\sum_i x_{ij}$ must equal the fixed amount of commodity j in the whole society for all j's, and no initial endowment for an individual i is strictly greater than the corresponding vector $(x_{i1}, x_{i2}, \ldots, x_{im})$. The market mechanism does not give a complete preordering of all those alternatives; it only singles out certain alternatives, namely, equilibrium points. Hence, every ordering that distinguishes equilibrium points from nonequilibrium allocations is compatible with the market mechanism

We will distinguish three sets. Denote by Q the set of equilibrium points, by P the set of Pareto-optimal points not in Q, and by S the set of attainable allocations not in P and not in Q, that is, the set of all non-Pareto-optimal allocations[3]

From this definition it follows that the union of Q, P, and S is the set of all alternatives A, $Q \cup P \cup S = A$ Hence, if all elements of the sets Q, P, and S are comparable by a relation, all attainable alternatives are comparable by that relation. We will define now such a relation The relation W is a preference relation, which gives the following ordering of alternatives each element of Q is strictly preferred to each element of both P and S, each element of P is strictly preferred to each element of S, and elements of the same set are indifferent[4]

It is easy to see that the relation W does not violate the requirements put on a preference relation· reflexitivity, transitivity, and completeness Hence, we have defined a preference relation compatible with the market mechanism We will take this relation W now as a social preference pattern and look at the market mechanism as a social welfare function, which gives us in cooperation with the relation W a preference ordering of all alternatives, that is, of all attainable allocations.

3. Positive Association

It is a precondition of Arrow's condition 2 (Arrow 1963), the condition of positive association of social and individual values, that it is possible that preferences concerning solely pairs of alternatives including a fixed alternative x can change with preferences concerning all other alternatives remaining constant. It can be proved rigorously that this precondition cannot be fulfilled in our case.

---

[2] Part (2) is implied if local nonsatiation (Arrow and Hahn 1971) is assumed for each individual, and it assures that the set of alternatives changes if the initial endowments change. Otherwise the social preference ordering could change without any change in individual preferences, because a change of the initial endowments will in general change the competitive equilibria Part (2) reflects that a different distribution of the society's endowments among the individuals implies a different set of alternatives The necessity of part (2) has become obvious to the author because of a critique by an unknown referee.

[3] It is not really necessary to distinguish between the sets P and S in this context, but it assures the Pareto-optimality of the welfare function defined in the following

[4] The case a preferred to b is denoted by $a \gtrsim b$, a strictly preferred to b means $a \gtrsim b$ and not $b \gtrsim a$ and is denoted by $a > b$; a indifferent to b means $a \gtrsim b$ and $b \gtrsim a$ and is denoted by $a \sim b$; a subscript k refers to individual k.

If there are two commodity vectors x^i and y^i and an individual i which used to prefer y^i to x^i and prefers now x^i to y^i, then there must exist a commodity vector z^i sufficiently close to x^i, such that the individual prefers y to z^i if it prefers y^i to x^i and prefers z^i to y^i if it prefers x^i to y^i. Since external effects are excluded, each individual orders allocations by ordering his commodity vectors of each allocation.

Obviously we have to alter Arrow's condition 2. We will state condition 2′, a condition of similar content, and show that it is violated by our welfare function.

Condition 2′

i) For given individual preferences, the society prefers strictly x to y. If individual preferences change the way that some individuals, who used to prefer strictly y to x, prefer now strictly x to y, with all other individuals having unchanged preferences, the society must still prefer x to y, at least weakly, in its new preference pattern.

ii) For a given set of individual preference patterns, the society is indifferent between x and y. Individual preferences change now such that some individuals, who used to prefer y to x, prefer now x to y, and for every alternative z different from y which those individuals ordered $z \gtrsim y$, they order now $z > y$. All other individuals do not change their preferences. Then the new social preferences, belonging to the altered set of individual preferences, must not prefer y to x; hence, it must order $x \gtrsim y$.

Perhaps an explanation of part (ii) will be useful. Contrary to (i), we start with x and y indifferent to each other for the society. But for y in individual preferences we demand more than in (i): if it falls in comparison to x, it must fall relative to every possible other alternative z, that is, there must not exist a z which falls in comparison to y.

It will be proved by examples that both (i) and (ii) are violated by the welfare function, defined above.

4. The Counterexamples

Let x be an equilibrium point for a given set of utility functions I and given endowments. Let y be an attainable allocation which is not an equilibrium point at utility functions I. We will show that it is possible that x ceases to be an equilibrium point and y becomes one, if preferences of some individuals change from

$$y \underset{k}{>} x \text{ to } x \underset{k}{>} y$$

with preferences of all other individuals remaining the same. Using the social preference ordering, induced by the market and defined in the above way, this means that a rise of an alternative x in some individual preference patterns with other individuals' preferences remaining constant leads to a change in the social preference from $x > y$ to $y > x$, hence, condition 2′ is violated.

There are two commodities a and b and three groups of individuals Each group has the same size and identical utility functions Groups 2 and 3 do not alter their preferences, whereas group 1 does change its preferences from

$$y \underset{k}{>} x \text{ to } x \underset{k}{>} y$$

Because of this change y becomes an equilibrium point, which is the paradoxon This is shown in the graphical representation on figure 1 First we must do the computations in order to show that no individual violates its budget constraint and that the sum of the commodities equals the sum of the initial endowments, which must be the case because there is no production

Let the initial endowments be $w_i = (4, 4)$ for all individuals, the prices

$$p^x = \binom{2}{1}, \quad \text{and} \quad p^y = \binom{1}{2} \qquad \text{for } x \text{ and } y \text{, respectively}$$

The social endowment is $\sum_i w_i = (12, 12)$ if we assume that there is one individual in each group Further,

$$w_i p^x = (4\ 4)\binom{2}{1} = 12 \quad \text{and} \quad w_i p^y = (4\ 4)\binom{1}{2} = 12$$

EQUILIBRIA	QUANTITIES					
	a_1	a_2	a_3	b_1	b_2	b_3
x	3 5	3 5	5	5	5	2
y	5	2	5	3 5	5	3 5

$$(1){:}\ (3.5\ 5)\binom{2}{1} = 7 + 5 = 12 \qquad (5\ 3.5)\binom{1}{2} = 5 + 7 = 12,$$

$$(2)\ \ (3\ 5\ 5)\binom{2}{1} = 7 + 5 = 12 \qquad (2\ 5)\binom{1}{2} = 2 + 10 = 12,$$

$$(3)\cdot\ (5\ 2)\binom{2}{1} = 10 + 2 = 12 \qquad (5\ 3\ 5)\binom{1}{2} = 5 + 7 = 12,$$

$$(x).\ \sum_i (a_i b_i) = (3.5 + 3.5 + 5, 5 + 5 + 2) = (12, 12) = \sum_i w_i,$$

$$(y){:}\ \sum_i (a_i b_i) = (5 + 2 + 5, 3.5 + 5 + 3.5 = (12, 12) = \sum_i w_i$$

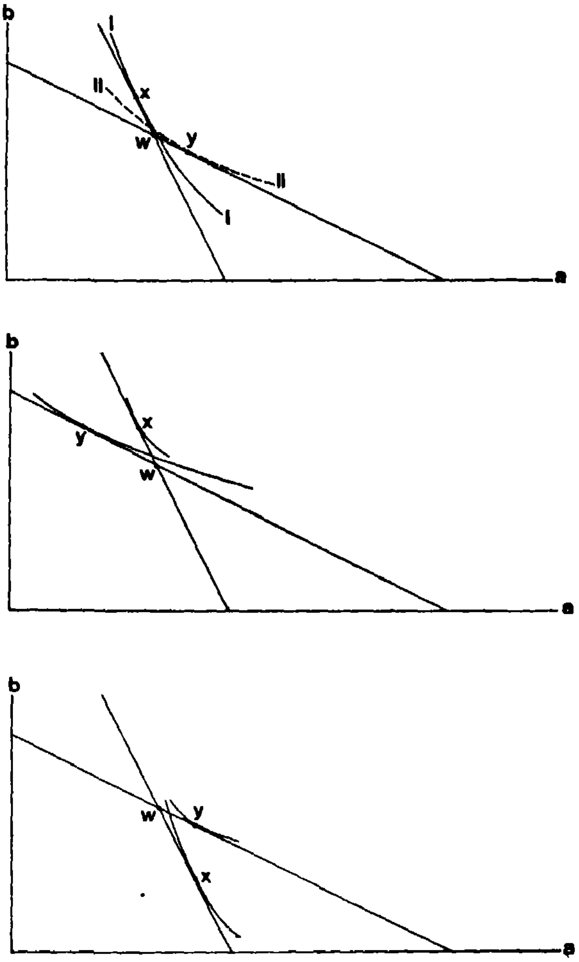


Fig 1 —*Top*, group 1 (individuals k) *Middle*, group 2 (individuals h) *Bottom*, group 3 (individuals l)

Hence, x and y are equilibria, if the indifference curves touch the budget lines in the allocation points of x, respectively y. Denote by

$$\left.\frac{\partial u^i/\partial a}{\partial u^i/\partial b}\right|^{j}_{\substack{a=\bar{a}\\ b=\bar{b}}} \qquad j = I, II$$

the ratio of the marginal utilities of the commodities a and b (hence the marginal rate of substitution) for individual i at its endowment $a = \bar{a}$, $b = \bar{b}$ and preference pattern j. Then in the example the following holds

For group 1 (individuals k):

$$\begin{aligned} 2 &= \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{I}_{\substack{a=3.5\\ b=5}} &&\neq \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{II}_{\substack{a=3.5\\ b=5}} \\ \tfrac{1}{2} &= \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{II}_{\substack{a=5\\ b=3.5}} &&\neq \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{I}_{\substack{a=5\\ b=3.5}} \end{aligned} \tag{1}$$

For group 2 (individuals h):

$$2 = \left.\frac{\partial u^h/\partial a}{\partial u^h/\partial b}\right|^{j}_{\substack{a=3.5\\ b=5}} \qquad \tfrac{1}{2} = \left.\frac{\partial u^h/\partial a}{\partial u^h/\partial b}\right|^{j}_{\substack{a=2\\ b=5}} \qquad j = I, II \tag{2}$$

For group 3 (individuals l):

$$2 = \left.\frac{\partial u^l/\partial a}{\partial u^l/\partial b}\right|^{j}_{\substack{a=5\\ b=2}} \qquad \tfrac{1}{2} = \left.\frac{\partial u^l/\partial a}{\partial u^l/\partial b}\right|^{j}_{\substack{a=5\\ b=3.5}} \qquad j = I, II. \tag{3}$$

Since at allocation x at preferences II and at allocation y at preferences I the marginal rate of substitution of individuals k differs from that rate of individuals not k (hence, individuals h or l), x is not Pareto-optimal at preferences II, and y is not Pareto-optimal at preferences I

Therefore, there cannot exist prices which make those points x, respectively y to equilibrium points at preferences II, respectively I, because Pareto-optimality is a necessary condition for an equilibrium point

We still have to show that part (ii) of condition 2′ is violated by our welfare function We can use the same example, leaving everything unchanged for group 2 and group 3 Preferences of group 1 must now be such that y is as well as x an equilibrium point at preferences I, hence, the society is indifferent between x and y If the new indifference curve through y, implied by preference pattern II, touches the old one in y and lies nearer to the origin in every point else for each z not y, the following holds:

$$z \underset{k}{\gtrsim} a$$

in preference pattern I implies

$$z \underset{k}{>} y$$

in preference pattern II. Obviously it is possible that two strictly convex curves touch each other in a point Since in the example everything except the indifference curves of group 1 is the same as in the previous example, only the graphical representation of group 1 is given Obviously part (ii) of condition 2′ is violated.

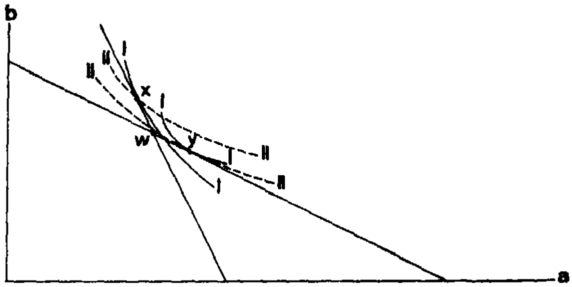


FIG 2

The dotted curves in figure 2 are the indifference curves at preferences *II* At preferences *I* $y \underset{k}{>} x$, and at preferences *II* $x \underset{k}{>} y$

$$2 = \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{I}_{\substack{a=3\ 5\\ b=5}} \neq \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{II}_{\substack{a=3\ 5\\ b=5}} \tag{4}$$

Hence, there cannot exist prices making x to an equilibrium point at preferences *II* because the marginal substitution rate between a and b is equal to 2 for individuals not k Therefore, x cannot.be Pareto-optimal. Hence, it cannot be an equilibrium point

Since

$$\tfrac{1}{2} = \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{I}_{\substack{a=5\\ b=3\ 5}} = \left.\frac{\partial u^k/\partial a}{\partial u^k/\partial b}\right|^{II}_{\substack{a=5\\ b=3\ 5}} = \frac{p^y_a}{p^y_b}, \tag{5}$$

then y is still an equilibrium point, and the social welfare function orders $y > x$ for the preference profile *II*, which violates condition 2′ part (ii)

5. Conclusions

The results reached show that the market mechanism has unsatisfying properties even under ideal circumstances[5] Those properties are immanent in the theoretical model, no matter if this model is an acceptable approximation of reality or not

It seems to me that the source of the violation of condition 2′ is that, for competitive equilibria besides the initial endowments, only the slopes of the indifference curves are important This means that primarily only the ranking of nearby alternatives matters, and the ordering of alternatives not being in a marginal neighborhood of each other comes in only by the necessity of transitivity and of the usual shape of the indifference curves [6]

---

[5] The confinement to a barter economy without production is no limitation to the validity of the results because of Rader's theorem given in n. 1

References

Arrow, K. J. *Social Choice and Individual Values* Cowles Foundation Monograph 12 2d ed. New York· Wiley, 1963.

Arrow, K. J, and Hahn, F. H. *General Competitive Analysis.* Mathematical Economics Texts, vol. 6. San Francisco· Holden-Day, 1971.

Debreu, G *Theory of Value* New York. Wiley, 1959.

Rader, J T. "Edgeworth Exchanges in General Economic Equilibrium." *Yale Econ. Essays*, vol. 4, no 1 (1964).

---

[6] Graphically two conditions are necessary and sufficient for an attainable allocation to be a competitive equilibrium the price line (the hyperplane in general) must (*a*) touch an indifference curve and (*b*) pass through the point representing the initial endowment, thereby (*a*) and (*b*) must hold for each individual Hence, with fixed initial endowments, only the slopes of the indifference curves at an attainable allocation x are relevant for the question, if x is a competitive equilibrium or not If we compare two situations differing only in the individual preferences (but not in the initial endowments), this means that it is relevant how preferences in a very small neighborhood of x differ in the two situations Of course, given the change of the preferences in all other areas of the positive orthant (that is, the slope of the indifference curves), the necessity of transitivity (the indifference curves must not cut) and the necessity of the usual shape of the indifference curves (convex to the origin) exclude certain possibilities of the slope of the indifference curves at x. The same is true for the question if a point y stays or does not stay an attainable allocation not being a competitive equilibrium if preferences change. The influence of the two necessities of transitivity and shape, respectively, diminishes in our example if x and y are "far from each other " Hence, if x and y are not "too close to each other," it is possible to change the indifference curves (without violating the necessities) in such a way that (*a*) $x > y$ change to $y > x$ and (*b*) the slopes of the indifference curves are accordingly Perhaps it is not too farfetched to point out an analogy to optimizing theory for local optima (as opposed to global optima) only an infinitely small neighborhood of the point in question is relevant

Miscellany

Drug-Law Enforcement: Should We Arrest Pushers or Users?

Any society wishing to reduce the consumption of an "undesirable" item, such as illegal drugs, has available two general strategies· reducing demand, by harassing buyers, or reducing supply, by harassing sellers Law enforcement agencies in the United States, presumably with the approval of the voting public, seem to have opted largely for the latter strategy ("we're not really interested in users, we want the pushers"), and yet ex-addict William S Burroughs (1966, p xl) advises. "If we wish to annihilate the junk pyramid, we must start with the bottom of the pyramid *the Addict in the Street,* and stop tilting quixotically for the 'higher ups' so called, all of whom are immediately replaceable" (italics in original) This note will consider what is involved in choosing between these two strategies and will argue that the United States has chosen to emphasize the strategy which, for a given enforcement expenditure, (1) probably has the smaller effect upon drug consumption, (2) seems likely to increase crime rates by addicts, making enforcement of other laws more difficult, and (3) seems more likely to corrupt police That something is wrong with our strategy seems clear enough The Government Accounting Office recently concluded from a year-long investigation that traffic in heroin is too big to stop at United States borders (*New York Times*, December 13, 1972, p 31) In some ways even more disturbing, however, was the recent announcement that about one-fifth of the heroin and cocaine seized in New York City over the past 12 years has been stolen from the police department (*New York Times*, February 1, 1973, p 1)

The economics of drug law enforcement is similar to that of sales taxes As is well known, who "pays" such a tax and its effects upon price and quantity depend upon the slopes of the demand-and-supply curves as well as upon the content of the law The law determines whether the buyer or the seller shall act as the state's "collection agent," in the case of sales taxes. Since punishment for violating drug laws is usually at least partly nonpecuniary, taking the form of jail time, social disgrace, and other inconveniences, the content of drug laws, together with police enforcement decisions, determines upon whom these nonpecuniary costs shall fall as well as upon whom the initial burden of pecuniary "taxes" (the expected value of fines, legal fees, etc.) shall fall If sale is prohibited and purchase not (by law, or de facto, by police decisions within the constraints of limited enforcement resources), the tax will fall initially upon sellers, who will convert the resulting pecuniary and nonpecuniary costs into a pecuniary increase in supply price to buyers Likewise, buyers will convert pecuniary and nonpecuniary inconveniences arising from police harassment into a pecuniary decline in demand price to sellers As is true with sales taxes, a given shift in demand relative to supply will result in the same effect upon consumption, regardless of whether the shift is in demand or supply Presumably, sales taxes are usually levied upon sellers because of lower enforcement and collection costs, a given amount of enforcement effort will produce more net revenue if directed at sellers. With sales taxes, we would wish to arrange collections so as to shift demand relative to supply as little as possible for a given tax rate

With drug laws, on the other hand, the goal is not to collect revenue but to discourage consumption, and we would wish to arrange laws and enforcement procedures so as to shift demand relative to supply as much as possible Whether we would wish to impose the "tax" initially upon sellers or buyers depends upon which arrangement we would expect to result in a greater shift in demand relative to supply This, in turn, would seem to depend mostly upon whether buyers or sellers are more efficient in converting the inconveniences of dealing with fines or nonpecuniary punishments such as jail sentences, social disgrace, etc , into pecuniary shifts in supply or demand The more efficient the conversion, the smaller the shift a given amount of harassment will produce It seems likely that sellers would have the advantage over buyers in making this conversion, since opportunities for specialization should attract to the drug-selling trade individuals who are peculiarly suited for it If the law requires that sellers occasionally cope with fines, jail sentences, and resulting social disapproval, we should expect successful drug dealers to be those who have the lowest supply prices for these services or are most adept at avoiding them Presumably, the successful practitioners of any trade are largely distinguished by their ability to convert pecuniary and nonpecuniary costs into small pecuniary prices, there is no obvious reason why drug dealers should be an exception In addition, specialization by drug dealers makes investment in devices for reducing the impact of enforcement (knowledge of whom to bribe to avoid arrest, which lawyers to hire to avoid conviction, etc.) more remunerative Finally, possibilities for entering the drug-selling trade will tend to make seller harassment less effective. A buyer jailed means one buyer less for the duration of the sentence; a seller jailed may simply mean fresh opportunity for a hopeful entrant.

Empirical evidence on the effects upon consumption of buyer harassment versus seller harassment is extremely difficult to obtain, but the meager evidence available seems to indicate that decreases in the effective penalty on buyers does lead to marked increases in consumption Wheat (1972) found that in Boston the expected costs of arrest to a user (defined as probability of arrest times probability of being sentenced to prison times average prison sentence) declined sharply between 1961 and 1970, while the estimated number of addicts in Boston increased by a factor of about 10 During the British experiment with decriminalizing heroin use and allowing physicians to prescribe heroin as they wished, addiction increased so greatly that the practice was abandoned in 1968

A second point to consider in choosing a drug-law enforcement strategy is the effect upon the pecuniary drug price Clearly, reducing supply will raise the price, reducing demand will lower it. In effect, reducing supply imposes an income tax upon those with highly inelastic demand ("addicts"?) Simon Rottenberg (1968, p 87 n.) claims that the price of a gram of heroin rises from $0 05 in Turkey to $295 retail in Chicago Since much of this price increase is due to enforcement of drug laws, the implicit income tax upon addicts resulting from seller harassment is substantial Adding to the troubles of these unfortunates is a perhaps regrettable concomitant of reducing drug consumption However, much of the income tax imposed upon addicts by current enforcement strategy seems to be shifted onto others by theft and other addict crimes Estimates of the proportion of property crimes committed by addicts range from about one-fourth to two-thirds (Erickson 1969, p 485, Wilson, Moore, and Wheat 1972, p 12) This heavy tax upon innocent bystanders has no obvious social function and could be largely avoided by concentrating enforcement effort upon users, reducing demand and the pecuniary price of drugs In effect, by shifting to such a policy we would be exchanging the current pecuniary tax upon addicts, which they easily shift onto innocent bystanders, for a largely nonpecuniary tax, which they could not easily shift onto anyone except drug sellers

Becker (1968, pp 193–98) has argued convincingly in favor of fines for criminal offenses whenever feasible Indeed, given the current strategy of concentrating on sellers, it would seem that a given amount of enforcement cost would have more effect upon drug traffic if expenditure currently used to maintain in prison those already arrested and convicted were diverted to making more frequent arrests or obtaining more convictions once arrests are made However, fines for drug users would seem as likely to increase property crime rates as the current seller-harassment strategy, since they result in an increase in the pecuniary price of drugs to addicts Nonpecuniary punishment of users is not so easy to shift onto others and also has the advantage of giving the community a mechanism for encouraging users to seek rehabilitation. Some authorities have claimed that many addicts will not seek rehabilitation without some measure of compulsion (Wilson et al 1972, pp 20, 25, 27, see also statements by Dr Herbert D Kleber of the Connecticut Mental Health Center reported by the *New York Times*, April 27, 1973, p 39)

Perhaps the most serious flaw of a policy of harassing drug sellers is that it lays the basis for a cooperative relationship between drug sellers and the police The gains from collusion in an industry are known to be greater the smaller the elasticity of demand, the greater the elasticity of marginal costs, and the smaller the costs of enforcing the collusion (Becker 1968, pp 205–7). It is widely believed that the elasticity of demand for hard drugs is very small (at least in the short run) The elasticity of marginal costs is presumably very large Thus, the gains from collusion would be large if colluding sellers could obtain cheap means of enforcing collusion and of preventing entry into the industry Enforcement policies directed against sellers furnish just that Schelling (1967, pp 119, 124–25) cited evidence that the Miami wire service syndicate "relied heavily upon the police as their favorite instrument of intimidation " Drug laws directed against sellers make this instrument available to colluding sellers for use against price cutters and potential entrants One way to view police policies of harassing sellers is as establishing a nonpecuniary "license fee " Those who are relatively efficient in converting arrest and other harassment into pecuniary terms will "pay the fee" and sell, others will be barred from the trade The sellers who succeed have an interest in enforcement of the law to prevent entry by others Both sellers and enforcement officials, however, have an interest in converting the license fee into terms other than harassment and arrest One possibility is to convert it into pecuniary (and illegal) bribes for police inaction, which raises ominous possibilities as to the types of people who will seek police work in the long run But colluding sellers may purchase police inaction by furnishing information on other criminal activity or by rendering other quite legal services to the police In this way, the police could avoid illegal and antisocial behavior Indeed, they could simultaneously serve their social mandate of reducing drug consumption and serve the interests of the colluding sellers by preventing individual members of the collusion from increasing their sales at the expense of other members and preventing new entrants from competing with established sellers.

Thus, enforcement strategy directed against drug sellers tends to enhance a natural accord between illegal business firms and the police based upon their common interest in civil order After visiting Harlem in May 1968, Harry McPherson (1972, p. 371) reported hearing of a conversation between Rap Brown and a Harlem rackets boss Brown had made a strong speech The boss told him: "I agree with a lot of what you said Except I don't want any riots I've got to raise $60,000 to buy off some people downtown on a narcotics rap. I can't do that if there's a riot. You start a riot and I'll kill you " Brown is rumored to have left town the next day.

References

Becker, G. S. "Crime and Punishment: An Economic Approach." *J P.E.* 76 (March–April 1968): 169–217.

Burroughs, W S *Naked Lunch* New York· Grove, 1966

Erickson, E. "The Social Costs of the Discovery and Suppression of the Clandestine Distribution of Heroin." *J.P.E* 77 (July–August 1969) 484–86.

McPherson, H. *A Political Education*. Boston· Little, Brown, 1972.

Rottenberg, S."The Clandestine Distribution of Heroin, Its Discovery and Suppression." *J P E.* 76 (January–February 1968) 78–90.

Schelling, T C "Economic Analysis and Organized Crime " In President's Commission on Law Enforcement and Administration of Justice, *Organized Crime* Task Force Reports Washington. Government Printing Office, 1967

Wheat, I D., Jr. "Heroin Abuse in Boston. The Dynamics of the Heroin Market and the Response of the Criminal Justice System." Mimeographed. Harvard Univ , John F. Kennedy School of Government, Public Policy Program, 1972

Wilson, J Q , Moore, M H , and Wheat, I D , Jr "The Problem of Heroin " *Public Interest* 29 (Fall 1972) 3–28

BILLY J EATHERLY

Mississippi State University

Puzzles and Problems

EDITORS' NOTE: This section, which is being terminated, has been edited by Stanley Fischer, to whom the editors wish to express their gratitude.

*Q*725—Donald A Walker, Indiana University of Pennsylvania Joan Robinson presented, in *The Economics of Imperfect Competition* (London Macmillan, 1933, chap. 26), the accepted account of exploitation of workers that occurs because of the relation between the marginal revenue product of labor (*MRP*) and the wage rate in monopsonistic labor markets, and the related matter of the magnitude of monopsony profits By constrast with her treatment, it is here (fig 1) assumed that a monopsony faces a supply-of-labor (*S*) curve that bends back toward the wage-rate axis in the relevant range It is assumed, for simplicity, that labor is the only factor used by the firm, and that the *S* curve and the *MRP* curve each have constant negative slopes It is also assumed that the average revenue product (*ARP*) curve cuts the supply-of-labor curve from above. In what ways does recognition of this situation lead to modifications of Robinson's analysis and conclusions?

Answer—It can be seen that Robinson's solution is not true in general She considers only the case of a positively inclined *S* curve Accordingly, she identifies, first, the intersection at *N* of the positively inclined marginal cost of labor (MC_L) curve and the *MRP* curve, second, the equilibrium volume of employment L_0, and third, the relevant point *H* on the *S* curve She concludes that the workers in a monopsony are exploited because the wage rate OW_0 is less than the marginal revenue product *OM* Under the conditions of the question, however, the MC_L curve lies below the backward-bending *S* curve, so that although the firm still maximizes profit by equating the MC_L to the *MRP*, the resulting wage rate OW_1 determined by the relevant point *V* on the *S* curve exceeds the marginal revenue product *OM*. It also follows that Robinson does not present a general treatment of the magnitude of monopsony profits. With the rising *S* curve of her analysis, the wage bill is OL_0HW_0, so monopsony profits are W_0HQA. By contrast, if the *S* curve is backward-bending, the wage bill is OL_0VW_1, so profits are the smaller magnitude W_1VQA.

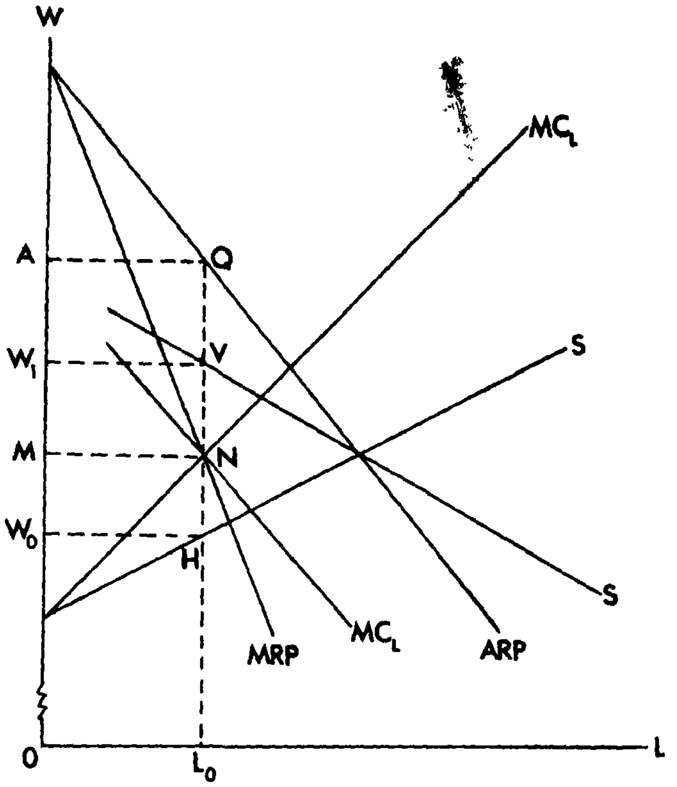


FIG 1

A comparison of the cases shows that there are two possible reasons for monopsony profits Some profits may be obtained by "exploitation" of the workers This will occur if the *S* curve is rising But exploitation is obviously not a source of profits if the *S* curve is backward-bending. On the contrary, in that case the workers are paid the sum $MNVW_1$ in addition to their contribution to revenue, at the expense of what would otherwise be a part of profits Even in the latter case, provided that revenue exceeds cost, it is advantageous for the firm to be a monopsonist, because the firm's market position enables it to obtain the profits shown by W_1VQA, as a permanent flow of income There is, therefore, a second reason for monopsony profits, one from which the other firm also benefits, namely, the absence of the freedom of entry and competitive equilibrating forces that would lead to a systematic equalization of revenue and cost.

Book Reviews

Monetary Theory: Inflation, Interest, and Growth in the World Economy. By ROBERT A MUNDELL.
Pacific Palisades, Calif.: Goodyear Publishing Co, 1971. Pp ix + 189. $10.00

During the 1960s, Robert Mundell wrote a series of short articles which had a sharp impact on the development of the monetary theory of both a closed and open economy. In the former category (represented by part 1 of this book) were Mundell's seminal *J.P E.* papers on inflation and on growth, in the latter (represented by part 2) were his important contributions toward the development of a general-equilibrium analysis of the monetary aspects of international trade.

In the first four chapters of part 1 of this book (and it is on this part that this review will concentrate) Mundell has conveniently collected the aforementioned *J P.E.* papers (somewhat modified)—and to them he has added two chapters (5–6) designed to integrate and generalize their respective analyses Part 2 is an extension of some of the basic analysis that appeared in his *International Economics* (1968) and consists primarily of new papers.

The point of departure of Mundell's analysis in part 1 is Metzler's famous diagrammatical representation of the general-equilibrium determination of the rate of interest and price level in a full-employment economy With the aid of this diagram Metzler made a fruitful distinction between the impact of a monetary change generated by deficit financing (which is neutral in its effects) and that generated by an open-market purchase (which, in Metzler's analysis, is not). In his first chapter, Mundell very insightfully demonstrates that this difference is not inherent in the open-market operation as such, but depends on whether or not the taxes used to finance the interest payment on government debt are capitalized by the taxpayers.

From here Mundell goes on (in chaps. 2–4) to make some ingenious modifications of the Metzlerian diagram whose common theme is that since the vertical axis of this diagram represents a percentage rate per unit of time (namely, the rate of interest), it can also be used to represent various relevant rates of change in the economy Here Mundell is at his best in most neatly generalizing the Metzlerian apparatus to the case of a continuous inflationary (or deflationary) process and in demonstrating the nonneutrality of such a process Mundell is also at pains to provide an economic interpretation of this conclusion in terms of the discrepancy created by such a process between the nominal and real rates of interest—and hence its differential impact on the demand for real money balances as contrasted with that for commodities. In a similar way Mundell analyzes the impact of growth on the economy in terms of the rate of change of prices that such growth generates, given the nominal supply of money.

All of us who work on monetary theory are in Mundell's debt for his incisive analyses of these problems. I must, however, admit that the new essays (chaps 5–6) in which Mundell attempts to elaborate upon this analysis and integrate it into the theory of growth are less helpful. I find the argument here difficult to follow, including the difficulty of seeing precisely how the earlier analysis is integrated into these essays.

Thus, the two equilibrium conditions used to carry out the analysis in the earlier chapters are the familiar IS-LM curves, representing, respectively, flow equilibrium in the commodity market and stock equilibrium in the money (or financial-assets) market. In the growth chapters, on the other hand, the two conditions relate only to the stock equilibria—namely, those for real-money balances and physical-capital goods, respectively—with an inadequate explanation of the way in which the commodity flow-equilibrium condition manifests itself in the stock-adjustment process Again, there is an inadequate explanation of the assumption that money is a factor of production—which plays a crucial role in this analysis. Nor, correspondingly, is there an indication of the way the theory could be developed under the assumption that money is only a consumption good.

Another point that should be noted is that strictly speaking Mundell is not dealing with a growth model, but with the theory of the accumulation of wealth to a desired (and thenceforth stationary) level. A crucial question which then arises in the present context is the influence on this desired level of wealth—and of the desired level of its physical-capital component, in particular—of an inflationary process Unfortunately, Mundell does not deal explicity with this question In any event, there must be something in the nature of his assumptions here that leads him to the conclusion that a higher rate of inflation necessarily leads to a lower long-run equilibrium level of real-money balances, whereas in other analyses of this problem the sign is ambiguous (cf Levhari and Patinkin 1968, sec. 11) This is a question on which I would like to have had Mundell's view

Another major theme of part 1 of Mundell's book is the effect of inflation in stimulating the rate of growth. Mundell analyzes this problem under different assumptions and to my mind does not probe deeply enough into the reasons these differing assumptions lead to different conclusions—though a common conclusion of all these cases is that the maximum contribution of inflation to growth is reached at relatively low (by Latin American standards!) rates of inflation, and that this contribution is at best "small." Thus, under certain assumptions (chap. 6) Mundell concludes that this optimum rate is roughly 10 percent, which increases the growth rate "only" by about 0.5 percent

What I find missing in this discussion, however, is an indication of what constitutes a "large" increase in the rate of growth. After all, the analysis here assumes a full-employment economy with a constant technology whose sole source of more rapid growth is a more rapid accumulation of capital Under such restrictive assumptions it is hard to see how anything but an extremely drastic policy could significantly affect the growth rate

As a case in point, consider Mundell's analysis on pages 38 through 39. The model described here is a Harrod-Domar growth model (with a capital-output ratio of 2), supplemented by the quantity equation. Under certain assumptions about the other relevant economic magnitudes (namely, the reserve ratio of banks and the elasticity of the demand for money with respect to the rate of change of prices), Mundell shows that the growth rates generated by different rates of inflation are described by the first two columns of table 1 Now, let me first note that these calculations are made on the assumption that the banks' reserve ratio is 30 percent, which in Mundell's model means that though every additional dollar generates inflationary pressures, only 30 percent of them generate capital accumulation and hence growth. Thus, if we assume as an alternative (as Mundell does indeed assume in other contexts) that all of the expanded money supply is used initially to finance investment (which is analytically equivalent to assuming a

TABLE 1

DETERMINANTS OF THE GROWTH RATE

ANNUAL GROWTH RATE (PERCENTAGE) (1)	ANNUAL INFLATION RATE (PERCENTAGE)		SAVINGS RATIO OR RATE OF INCOME TAX (4)
	Reserve Ratio = 30% (2)	Reserve Ratio = 100% (3)	
0.00	0.0	0.0	0.0
0.25	5.7	1.3	0.5
0.50	14.25	2.8	1.0
0.75	28.5	4.4	1.5
1.00	57.0	6.25	2.0
1.25	142.5	8.3	2.5
1.50	∞	10.7	3.0
5.00		∞	10.0

reserve ratio of 100 percent), then the relevant rates of inflation are given by column (3) of table 1, which are most moderate as compared with those of column (2). I offer these alternative calculations not because they are necessarily more realistic than those based on Mundell's assumption, but only to illustrate the sensitivity of the calculations to changes in the assumptions.

But what is far more important for my present purposes is to note that if we delete the monetary part of Mundell's model, we are left with a Harrod-Domar model whose rate of growth (under Mundell's assumption of a capital-output ratio of 2) is half the savings ratio. In column (4) of the table are accordingly presented the growth rates associated with various savings ratios. Alternatively, if in the spirit of the Mundell model we assume that the private sector consumes all of its disposable income, and that investment in the economy is carried out entirely by government—but this time financed by an income tax—then column (4) also represents the income-tax rates necessary to generate different rates of growth.

And then we can ask the question, Which represents a "more drastic" policy—one which roughly doubles the rate of inflation in order to increase the rate of growth from 0.50 to 1.00 percent per year (as is shown by col. [3]), or one which—in order to accomplish the same objective—doubles the rate of income taxation (as shown by col. [4])? The criteria by which one would answer such a question are not clear to me.

As indicated at the beginning of this review, I have concentrated on part 1 of Mundell's book. Many readers, however, will be primarily interested in part 2, in which Mundell—again with great ingenuity—develops a graphical general-equilibrium analysis of the monetary aspects of international trade. Here he emphasizes the view that balance-of-payments adjustments (particularly those of a small country) should be analyzed in terms of the movement of international prices relative to domestic ones, and not primarily in terms of the movements of import prices relative to export prices. He also elegantly demonstrates—far more rigorously than has hitherto been done—the somewhat neglected truth that in a gold-standard world the money supply of any given country is endogenously determined. Correspondingly, he draws the implication of this fact for the money supply of a country which carries out a devaluation in a world of fixed exchange rates.

More generally, Mundell bases his analysis of the international financial system on the supply of—and demand for—international monetary reserves by the various countries of the world. He then uses this analytical framework to explain some of the dramatic changes that have taken place in this system since 1968.

These few words certainly do not do justice to the depth and ingenuity of Mundell's analysis in part 2. Let me then simply conclude by saying that Mundell's book must and will be carefully studied by every economist working in the fields of monetary and international-trade theory. Mundell's terse and elliptic style of writing does not make for easy reading, but the reader who overcomes these difficulties will find himself greatly rewarded by the additional insights into difficult problems that he thereby gains.

References

Levhari, David, and Patinkin, Don "The Role of Money in a Simple Growth Model." *A E R* 58 (September 1968) 713–53

Mundell, Robert A *International Economics*, New York Macmillan, 1968.

DON PATINKIN

The Hebrew University of Jerusalem

Theory of General Economic Equilibrium By TROUT RADER
New York and London Academic Press, 1972 Pp xx + 362 $14 50.

This is a companion volume to the author's *Theory of Microeconomics*, using the distinctive notation and terminology of the earlier work as well as many of the results derived there Taken together, these two volumes represent a truly impressive attempt to examine and extend the current state of knowledge in a number of major areas of modern analytical economics, with detailed consideration of the most important results in each area studied. The broad range of coverage of the volume under review is indicated by a partial listing of contents· intertemporal efficiency criteria, turnpike theorems, microgrowth models, existence and optimality of the competitive equilibrium, bargaining games, stability and comparative statics analysis of the competitive equilibrium, and applications to trade theory and macroeconomics In essentially all of the areas covered, known results are reformulated, extended, and generalized, and new proofs are constructed It should be emphasized that this book is not so much a textbook as it is a treatise summarizing a variety of Rader's recent research, much of it published here for the first time. If judged solely as a textbook, there are some serious flaws in this volume, but these are precisely the flaws one would expect in a creative work; in particular, the book tends to be unbalanced in the direction of work in which Rader has been engaged, and even there one finds occasional slips and omissions.

This is an exceedingly formidable book from a mathematical point of view, especially in the first chapters dealing with infinite horizons. Because of the organization of the book, the mathematical requirements are not so stringent in the middle and latter stages, but many readers may simply give up when faced with the problems of getting through chapter 1 (exercise 5, p. 13: prove the Bolzano-Weierstrass theorem). The book is also not completely self-contained; unless the reader first studies *Theory of Microeconomics* quite thoroughly, reading this volume tends to be frustrating, since so many of the concepts with which the reader is assumed to be familiar (directionally dense preferences, interindustry production, no free production in the limit, etc.) are discussed in detail mainly in the earlier volume. A glossary of terms and a recapitulation of proofs from the earlier volume used in this volume would be a great help. The method of approach tends to be one of formulating a general and rather abstract framework, following which the more conventional specialized models are examined. While this has some obvious advantages from a formal point of view, it leads to the choice of a notation that tends to be somewhat cumbersome when considering the simpler cases, and makes it more difficult to utilize the intuition one has gained from considering the conventional cases.

It really requires an economist with Trout Rader's range of interests to adequately judge the contribution in total of this volume; for the rest of us, we can at most attempt to evaluate specialized aspects of the book. For example, I tried to look in some detail at chapter 5 (Macrostatics) and found several slips, all of them, however, easily corrected. Theorem 11 (p. 210) is incorrect because the sum of diagonal elements is not constrained to be negative. Problem (iii), page 245, cannot be solved unless, for example, one specifies that the coefficient matrix is Metzler-Mosak, while in problem (i), page 248, the matrix is quasinegative definite (rather than indefinite), and the condition should hold wherever evaluated. The best-known result concerning "sometimes stable" matrices—page 205 (DA is stable for some diagonal D with positive diagonal elements), that is, that A is sometimes stable if A contains a nested sequence of principal minors, the ith of sign $(-1)^i$, $i = 1, \ldots, n$—is not mentioned. On page 211, "market laws have been used with substantive conclusions only for the gross substitutability case" neglects, for example, the case where the nonnumeraire goods obey the Morishima conditions. In contrast to these minor flaws, there is an elegant treatment of the derivation of the Hicksian laws and a particularly straightforward analysis of factor-price and consumption configurations in international trade also contained in chapter 5.

The most interesting aspect of this book to me is the work on pairwise optimality and the general attempt to incorporate the more general framework of n-person cooperative games into economic theory, following the work of Scarf, Debreu, and others. The sections on stabilization of markets and price and price-asset adjustment mechanisms also represent important areas for further research.

In summary, this book, if used as a text in an advanced theory seminar, should be used in conjunction with *Theory of Microeconomics*; it is accessible only to those with an extensive background in mathematics, and it occasionally contains erroneous or ambiguous results. But I think it would be ideal for such a seminar, and despite its faults, it represents an impressive accomplishment.

JAMES QUIRK

California Institute of Technology

Technology and American Economic Growth. By NATHAN ROSENBERG.
New York. Harper & Row, 1972. Pp. 211.

This book is really a long essay, or perhaps a series of shorter essays. As such, it does not present the results of new historical research. Rather, it integrates economic analysis with the results that have been obtained from existing historical studies (some of them done by Rosenberg himself) on the role of technology in the development of the American economy. The result of this integration is a comprehensive survey both of the causes, especially the economic causes, of technological change and of the effects of technological change on the American economy.

Rosenberg has done an excellent job in handling the task he set himself. He has achieved remarkable insight into the nature of the processes he describes and he is able to communicate his ideas clearly and concisely. I was particularly impressed by the chapter (chap. 2) which analyzes the influence of economic factors on the development and application of new technology and by the two chapters (chap. 3 and 4) which describe American experience in the nineteenth century. As is natural, given Rosenberg's own research interests, the discussion of the role played by the American machine tool industry is especially outstanding. The chapter on the twentieth century (chap 5) is also good, but I did not feel that it added as much to my understanding of events as did the earlier chapters.

In reading this book I must confess to feelings of both admiration and frustration. Certainly I wish I had written it Moreover I feel that I should have been able to write it, all the while realizing, however, that such an attempt on my part would have lacked much, if not most, of the considerable elegance and originality displayed by Rosenberg

In my judgment, almost any student (or professor) of economics or history, not to mention economic history, will profit from reading this book. More important, the benefit will almost certainly vastly exceed the opportunity cost of the relatively small amount of time required to read it. Indeed, my only important criticism of the book is a direct result of my concern that it be widely read, especially by persons without extensive training in economics. Right at the start (on p 4), there are two diagrams utilizing isoquants and price lines (without, however, explaining or even defining these concepts). I think this is unfortunate, because the diagrams are not needed by economists and they are more likely to confuse than to help readers innocent of economics What is even worse, by giving a false impression of the level of competence in economic theory needed to benefit from the rest of the book, the diagrams may well frighten off potential readers who, in fact, have much to gain from it

My other complaint about this excellent volume is directed at the publisher. There is no bibliography A book as rich in historical examples and references to the literature as this one is needs a bibliography The convenience to the reader, or at least some readers, would have been well worth the extra cost involved.

LARS G. SANDBERG

Ohio State University

Planning for Economic Growth in the Soviet Union, 1918–1932. By EUGENE ZALESKI. Translated from the French and edited by Marie-Christine MacAndrew and G. Warren Nutter.
Chapel Hill: University of North Carolina Press, 1971. Pp. xxxviii+425. $15 75.

This is one of the most important Western studies of Soviet economic planning. The heart of Zaleski's volume is a pioneer comparison of plans and revisions of plans with performance during the period of the First Five-Year Plan (1928–32). It embraces five-year plans, annual plans, and—for selected periods—quarterly and even monthly plans, and considers the progress and failure of Soviet planning both chronologically and analytically.

Zaleski shows that, in a certain sense, planning for economic growth took place successfully in the USSR in the period described. Even during the years of over-ambitious planning, 1930 and 1931, when targets were "dreams rather than plans," "all these false notions, more or less shared by the authorities, nevertheless enabled the plan to be turned into a powerful instrument of growth " This was because continuous pressure from a vast bureaucratic planning machine replaced the market in moving the economy in the direction of the aspirations of the authorities.

However, Zaleski convincingly shows that while the periodic business cycles characteristic of market economies were eliminated, they were replaced by new forms of economic fluctuation Whereas in market economies fluctuations tend to be transmitted rather uniformly, but with a lag, in the Soviet economy the fluctuation spreads "according to the strengths and weaknesses of the dams that the planning sector puts up in each sector " In prerevolutionary Russia all parts of industry were affected by the business cycle, but in the single period 1930–34 eight types of fluctuation can be distinguished in different industries. very rapid growth (e g , tractors), rapid growth with shock (machine tools), continuous growth (electricity), interrupted growth (steel), stationary state (cotton fabrics), stationary state followed by shock (cement), violent shock (sugar), and decline (meat and fat) All this is illustrated with a quite extraordinary wealth of statistical data, meticulously collected.

Some of Zaleski's conclusions may, however, need revision. He frequently compares growth rates in 1922/3–1927/8 with those in 1928/9–1932 and does not always take sufficient account of the unused capacity of the earlier period which was crucial, even in 1927/8, to the higher growth rates then achieved He concentrates his discussion of successes and failures in planning on performance in terms of growth rates, even though he rightly points out that the overriding criteria of success for the Soviet authorities were "the execution of large-scale projects, the achievement of social transformation and the build up slowly and sometimes empirically of a particular planning system." How far was the retardation in growth in 1930–33 due to deficiencies in planning, as Zaleski implies, and how far to the lack of spare capacity, coupled with the fact that these were years in which the new plants to be brought into operation during the Second Five-Year Plan were not yet completed? Zaleski's forthcoming volume, dealing with Soviet planning and growth after 1932, may help us to weigh these different factors.

The present volume is a translation of the French edition published in 1962, with some minor revisions (including a critique of Feldman's path-breaking but highly unrealistic growth model), and with a useful index and bibliography added. A fuller paginated table of contents would have helped the reader to use this complicated volume. When the next volume appears in French, it would be tempting to urge the very competent translator and editor to acquire temporarily some of the "shock" and "storming" characteristics possessed by Soviet planning in the 1930s, rather than keep us waiting 9 years for the English version, if it were not for the strong danger that the high quality of the present edition would not be maintained.

R. W. DAVIES

Centre for Russian and East European Studies
University of Birmingham

ECONOMICS IS A MATTER

ECONOMICS, Ninth Edition

Paul A. Samuelson, Massachusetts Institute of Technology. 1973, 917 pages, $11.50.

Changes in American life and economics and in attitudes towards the economic system provide the motif for the revised ninth edition. Professor Samuelson recognizes that to many of the younger generation, Gross National Product is really Gross National Pollution. Polluting, dehumanizing growth, merely for the sake of material growth, is something no one wants. Therefore in the ninth edition, Professor Samuelson has not abandoned the national income approach or the concept of GNP, but he has shown how the Gross National Product must now be supplemented by a life-quality indicator that he calls Net Economic Welfare, N.E.W. This concept is adapted from research done by William Nordhaus and James Tobin of Yale, and their estimates are the basis for the first figure in the text and for the discussion of National Income Accounting in Chapter Ten. The new Chapter 42, Winds of Change: The Evolution of Economic Doctrine, offers perhaps for the first time in an elementary economics course a section devoted to the history of the subject and its relationship to the history of ideas. The problems of the quality of life are covered in Chapter 40, and the cost-push inflation or the problems of the modern mixed economy are thoroughly discussed in Chapter 41.

BASIC ECONOMIC PRINCIPLES, Third Edition

Robert A. Lynn, Kansas State University. 1974, 416 pages (tent.), $9.95 (tent.). *Instructor's Manual.*

As a result of a survey conducted among users of the previous edition, Professor Lynn's third edition of this text has been tailored to meet the needs of its of readers. Briefer and simpler than before, the text covers a broad spectrum of macro- and microeconomics topics, using a public policy oriented approach. It now contains increased coverage and updating of such topics as communist economics, dollar devaluation, and direct government regulation of the economy.

McGRAW-HILL BOOK COMPANY

Private Savings: Ultrarationality, Aggregation, and "Denison's Law"

Paul A. David
Harvard University and Stanford University

John L. Scadding
Stanford University

The article examines the behavior of the ratio of gross private saving to gross national product—the gross private savings rate or GPSR—for the United States during the period 1898–1969 It emerges that the GPSR defined to include consumer durables expenditures and the imputed return on durables has no trend over the period and that its year-to-year variation is extremely small except for the World War I, World War II, and Great Depression years This stability prevailed in the face of pronounced shifts in the composition of private savings from personal saving to durables expenditure and corporate saving, and a redistribution of income from the private to public sectors. The notion of ultrarationality, in which households subsume corporate and government spending and saving in their budget decisions, is used to explain the stability of the GPSR despite the important changes in its composition. Leading implications of this stability for theory and policy are explored: one important conclusion is that, under full employment, fiscal policy can neither affect aggregate expenditure nor alter the ratio of investment to consumption

I

The most profound and lasting impact of the Keynesian revolution on economic thought has been to leave macroeconomics with a strong

The proximate origins of this paper lie in collaborative research undertaken by the authors in the summer of 1969 This effort is built upon work previously carried out in connection with the Stanford Project on Economic Growth in the United States. We gratefully acknowledge the Social Science Research Council's financial support of this work, as well as the continuing advice and encouragement we received from Moses Abramovitz Warren Sanderson and Fred Nold furnished able assistance in the preparation and statistical analysis of the data. In preparing the present version of this paper for publication we have had the benefit of comments received from Edward Denison, Robert J. Gordon, Joseph Pechman, Eytan Sheshinski, and an anonymous referee on an earlier much longer draft entitled "What You Always Wanted to Know about 'Denison's Law' But Were Afraid to Ask."

predisposition to emphasize demand over supply conditions. Nowhere is this more evident than in studies of the determinants of aggregate savings. Much of the evidence on savings behavior is not direct but, rather, implicit in the results of the many studies of the consumption demand This means that not only is savings treated as a residual but also its study is confined as a rule to the savings flows deriving from decisions about the disposition of personal income, a curious restriction given that the corporate and personal contributions to aggregate private savings in the United States have been roughly equal over the past 70 years.

In the exposition of simple income determination theory, it is convenient to treat all savings as done by the household sector and to regard all investment as taking place in the firm or corporate sector, that is, to neglect consumer durables expenditure and corporate savings Alternatively, this simplicity can be achieved by treating expenditure on durables as part of investment demand and aggregating corporate and personal saving by defining the saving function in terms of private saving and (in the absence of government) total income rather than in terms of personal saving and personal disposable income alone

The direction of research on the demand for consumer durables has already made quite respectable the first type of aggregation required by this alternative approach Formally, the theory of consumer durables expenditure is a part of modern investment theory even though in empirical applications it has been thought essential to distinguish among the different categories of investment demand But the notion that corporate and personal saving can profitably be treated within a unified conceptual framework, let alone usefully aggregated for empirical purposes, has made little headway among economists at large; here theory and practice are in accord, and disaggregation has been the order of the day

This resistance seems curious on at least two counts First, it is possible to refer to a simple theoretical construction which supports the aggregation of the various savings flows [1] The model in question views households as "ultrarational" in the sense of regarding the corporate sector as an extension of themselves, an instrument of their private interests, the distinction between corporate and household sectors therefore is a legal but not an economic one One intriguing implication of this view—which is so at odds with the tradition of Berle and Means—is that corporate and personal saving are perfect substitutes Hence, a savings function defined in terms of total private savings should turn out to be more stable than separate savings functions for the personal and corporate components, in other words, aggregation is desirable.

---

[1] Cf. Grunfeld and Griliches (1960) for development of the (parallel) argument that interdependence of firms' investment plans justifies aggregation across enterprises within the corporate sector.

The formal construction of a transsectoral savings model of this sort is due to Bailey (1962, 1972) Although in Bailey's work the focus is on the characterization of "rationality" in the relations between the private and public sectors, clearly the corporate and personal sectors could be substituted without any change other than an appropriate relabeling of variables [2] As we shall see, the Bailey formulation is a better description of relations within the private sector than it is of connections between the private and public sectors

The second ground for surprise at the slight attention paid the question of aggregating savings is that economists have had before them for some time intriguing empirical evidence that (gross) private savings is a very stable function of GNP, and that it has in fact been much more stable than either corporate or personal saving considered separately More than a decade has now passed since the publication of Denison's 1958 article calling attention to the remarkable stability of the ratio of gross private saving to gross national product in the United States during the postwar years 1948–56 Over that period the average level of the gross private savings rate (GPSR), as this ratio is called, just about equaled the GPSR for 1929, the last date before 1948 that Denison thought was comparable in the sense of being a full-employment, peacetime year

These two propositions—that the year-to-year variability in the GPSR was very small and that at least since 1929 there was no trend in the rate—were dubbed "Denison's law" by Hickman (1966), who appears to have been alone in responding to Denison's (1958, p 267) invitation to future students of the American economy to inquire whether this empirical regularity would persist beyond the period he had been able to observe

The purpose of this paper is to promote wider recognition of the reign of Denison's law by presenting both a theoretical framework for its interpretation and the results of more extensive empirical tests of its validity Denison's findings for the United States and our own are com-

---

[2] Much has been made in the literature of including imputed income on consumer durables in estimates of consumption and disposable income. But little attention has been paid to the question of whether corporate income and saving should be included in household budget constraints There are a few exceptions Friedman (1957, pp 116–23) explicitly raises this question twice, and Friend and Schor (1959), in their investigation of the distribution of saving by income class, impute corporate saving to the household sector But they are quick to point out that the procedure produces a departure "from any *normal* concept of individual's savings" (p 237, italics added) Another proposed modification of disposable income is the addition of capital gains. One argument for doing so is that capital gains reflect transfers from the corporate to the personal sector and hence "should be included at least in personal income and saving" (Bhatia 1971, p 364). A second reason is to capture changes in the stock of wealth when wealth is one of the arguments in the consumption function (Arak and Spiro 1971). Of course, it is possible to regard wealth, or at least a part of it, as a proxy for corporate income and saving, but typically it is included because it summarizes *future* disposable income. In other words, the aggregation implied in using wealth is over time rather than across sectors.

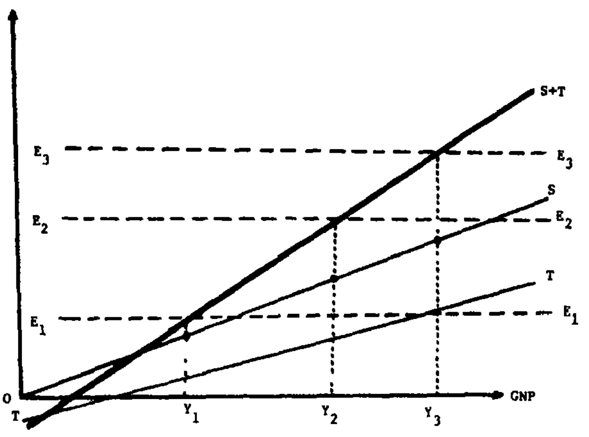


Fig 1

patible with a picture like that in figure 1 The ray *OS* through the origin depicts the observed stability in the (average) GPSR The curve *TT* described the dependence of tax revenue net of transfers, *T*, as a positive function of output The sum of the two curves, shown by the heavy line *S* + *T*, gives the amount of "leakages" from the system as a function of GNP To close the system and determine the equilibrium level of income, we add the schedule of "injections" shown by the curves *EE* The quantity *E* is the sum of gross private domestic investment, net foreign investment, and government purchases of goods and services

The intersection of *S* + *T* and *EE* determines the level of GNP Variations in *EE*, illustrated in figure 1, change the level of output, and the resulting levels of GPS and GNP trace out points on *OS* The level of consumption expenditure can be shown by adding the usual 45-degree line (not shown) through the origin the vertical distance between it and *S* + *T* gives the amount of consumption spending Note that with a stable GPSR, a change in the average tax rate is completely offset, for a given level of income, by a change in the average propensity to consume out of GNP in the opposite direction This means that unless the tax schedule is stable, consumption is not a stable function of GNP. Nor is it a stable function of disposable income The marginal propensity to consume is not independent of whether the source of the change in disposable income is a change in output or a change in the tax rate. But the loss is inconsequential, since, as figure 1 illustrates, a stable consumption function is not necessary for the simple Keynesian analysis of income determination The simplicity of income determination when the GPSR is stable can be more easily seen if *E* and *T* are combined as in figure 2*a*. Figure 2*a* shows clearly that the tax and expenditure multipliers are the same and that the common multiplier is just the inverse of the

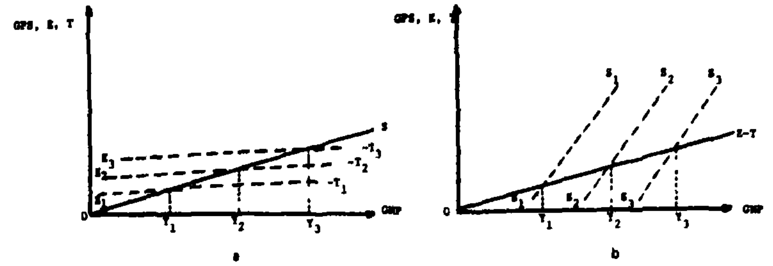


FIG. 2

GPSR. Figure 2*a* is not a perfect translation of figure 1. In the former we have drawn the *E* – *T* curve with a positive slope, which means that the corresponding *EE* curve must have a slope greater than that of the tax schedule, *TT*. Hence, in figure 2*a* the marginal propensity to spend exclusive of consumption spending is greater than the marginal tax rate, whereas in figure 1 this is not the case and its representation (in a diagram like figure 2*a*) would call for a downward-sloping *E* – *T* curve with respect to GNP.

We draw the *E* – *T* curve with a positive slope in figure 2*b* only to make the case for a stable GPS function no stronger than necessary. Thus, if *E* – *T* is *positively* related to GNP, an alternative explanation for the observed stability of the GPSR is that *E* – *T* is the more stable function of output and that the average propensity to spend minus the average tax rate is constant. Figure 2*b* illustrates this alternative, representing *E* – *T* by the ray through the origin, the constant GPSR is traced out by an unstable GPS function, denoted by *SS*, which wanders up and down the *E* – *T* ray. Note that figure 2*b* is not simply a relabeling of the schedules in figure 2*a*, because stability considerations require the slope of the *SS* function to exceed that of the *E* – *T* ray.

We regard this second explanation for the observed stability of the GPSR as the weaker of the two for the following reasons. First, it begs the question of why the amalgam of different types of expenditure and taxes should have a unitary elasticity with respect to GNP. Second, it entirely ignores the autonomous components of *E*, that is, those unrelated to GNP, whereas it is more conventional to treat much (if not all) of *E* as autonomous—in which case, *E* can hardly be a *stable* function of GNP. Finally, the marginal propensity to spend out of GNP must be fairly large if the *E* – *T* curve is to have a positive slope. For example, at current (third quarter, 1971) rates, the aggregate average tax rate is about .31 while the average propensity to consume out of GNP is .47. Assuming that average and marginal propensities are the same, the total propensity to spend out of GNP would have to approach an implausible .80 if the response of *E* – *T* to GNP is to be positive.

The next two sections of this paper demonstrate the validity of Denison's law for a much longer time span than in the earlier papers by Denison and Hickman On this point we are the beneficiaries of two developments which were not accessible to Denison (1) the generation of extra data points just by the passage of time since Denison wrote and (2) the fruits of historical research in estimating national accounts aggregates for the years before 1929. We use this wealth of additional data to show that with some slight modification Denison's propositions regarding the secular constancy in the GPSR, the remarkable degree of year-to-year stability in that rate and the far greater instability in the component personal and corporate saving rates are all borne out by the accumulated evidence of over 70 years

Section III advances some explanations for our findings by making use of the notion of ultrarationality mentioned earlier, and some further implications of this style of explanation are considered briefly in the concluding section, IV We should make it clear at this point that no pretense is made to provide a complete explanation, and our discussion based on the idea of ultrarationality is meant primarily to be provocative Our aim in this paper must be to confront economists with the *fact* of a stable GPSR, because the implications of that fact for both policy and accepted theory are too important to go on being ignored Denison originally challenged the profession to see whether his findings would stand up to additional evidence We maintain that in essence they do, and we leave in place of that challenge another one to provide a satisfactory explanation for the *why* of Denison's law

II

The regression equation for GPS presented by Denison for the period 1929–40/1948–56 is reported in table 1 The ΔGNP^* variable is the difference between the last "high employment" year GNP and the current GNP A "high-employment" year was designated by Denison as one in which the unemployment rate remained below 6 percent Plainly, the intention was to differentiate the mass unemployment experience of the thirties from the other peacetime years since 1929 As it turns out, Denison might equally well have taken the previous peak value of GNP, inclusive of the level attained in the current year In empirical experiments we have determined that in Denison's period it makes essentially no difference which definition is chosen

That being the case, we are free to interpret ΔGNP_t^* as the difference between the previous peak and current levels of aggregate income, and on such a reading the formulation of Denison's law in table 1 bears an obvious correspondence to the models of consumption behavior made familiar by Duesenberry (1949), Modigliani and Brumberg (1954),

TABLE I

DENISON RELATION FOR GROSS PRIVATE SAVING ESTIMATED FOR 1898–1916/1921–40/1948–64 AND SUBPERIODS CHOW TEST RESULTS

PERIOD	DEGREES OF FREEDOM	COEFFICIENTS ON GNP	COEFFICIENTS ON ΔGNP[d]	$\bar{R}^2$	D-W	STANDARD ERROR OF REGRESSION AS PROPORTION OF MEAN OF DEPENDENT VARIABLE	CHOW TEST FOR EQUALITY OF COEFFICIENTS BETWEEN SUBPERIODS
			Part A Between 1921–40 and 1948–64				
1929–40/1948–56							
Original Denison regression	20	1463	– 1237	[a]	[a]	[a]	$F(2,33) = 0.607$
Denison regression reestimated using revised data	20	1544	– 1339	997	1 059	043	$F_{.05}(2,33) = 3.29$
		(130 16)	(9 938)				
1921–40/1948–64	35	1552	– 1376	997	1 784	0 49	
		(161 25)	(7 407)				
1921–40	18	1492	– 1231	889	2 137	[b]	
		(35 44)	(7.417)				
1948–64	15	1554	– 7633	985	1 701	034	
		(123 50)	(0 362)[d]				
			Part B Between 1898–1916 and 1921–40/1948–64				
1898–1916/1921–40/1948–64	54	1533	– 1381	997	1 742	016	$F(2,52) = 16\ 123$
		(180 25)	(8 292)				
1898–1916	17	1835	– 6626	761	0 993	[b]	$F_{.05}(2,52) = 3.180$
		(24 428)	(2 53)				
1921–40/1948–64	35	1552	– 1376	997	1.784	049	
		(161 25)	(7 407)				

ADDENDUM

	1929/1948–56	1957–64	1957–69	1921–28[c]/1957–69
Standard error of forecasts from Denison relation as proportion of mean GPS		032	054	058
Standard deviation of statistical discrepancy as proportion of mean GPS	037	011	015	017

[a] Not reported
[b] Not calculated
[c] Full employment years 1925, 1926, and 1928 only
[d] Not significant at 95% confidence level

and Ando and Modigliani (1963) Alternatively, the previous peak GNP and ΔGNP_t^* may be read as the respective GNP counterparts of Friedman's (1957) concepts of a permanent component and a transitory component of personal income. Note that on the latter interpretation the estimates in the first equation of table 1 imply that the marginal propensity to save out of aggregate transitory income is greater than the corresponding propensity with respect to aggregate permanent income, a result quite in keeping with Friedman's findings regarding the direct savings behavior of the household sector One further virtue of making this switch to a "peak income" definition is that, in investigating the applicability of Denison's law to U S experience prior to 1929, it obviates the need to depend on the comparatively less reliable estimates of the rate of unemployment

The effect of the Commerce Department's 1965 revision of the national accounts was to raise the mean level of the GPSR for 1948–56 rather more markedly than the mean for the ensuing subperiod (1957–64) examined by Hickman; the revision for the first is from 1480 to 1543, while the revision for the second is only from 1527 to 1557. The result is that the stability of the GPSR within the 1948–64 interval appears still more striking in the revised data A reestimation of our version of Denison's equation for the period 1929–40/1948–56 using the revised Commerce Department figures is reported in table 1 The results there show that the only effect of the revisions of the national accounts has been to shift the coefficient attached to GNP, which is the estimate of the GPSR, upward from 1463 to .1544 Denison's specification implied a constant *average* GPSR for $\Delta GNP^* = 0$ This was tested by estimating a constant term, which turned out to be statistically insignificant at the 95 percent confidence level

The fit of the Denison relation remains spectacularly good, and, although the high $\bar{R}^2$ should be taken with a grain of salt in view of the serial correlation present in GPS and GNP, the closeness of the fit is equally manifested by the low (043) ratio of the standard error of estimate to the mean of the dependent variable

The latter statistic is also indicative of the extremely small year-to-year variability of the GPSR around its estimated "full employment" value Note that the ratio for the regression for 1948–64 alone is even smaller, .034 Proportional "errors" of this magnitude appear trivial indeed when ranged alongside the comparative size of the contemporaneous statistical discrepancy in the national income accounts. The standard deviation of the latter as a fraction of the mean of gross private savings was .037 for the years 1929/1948–56 (addendum in table 1).

The temporal scope of the empirical regularity isolated by Denison may now quickly be shown to be rather wider than the 1929–40/1948–56 results reported in the first two regressions of table 1. For, as the addendum of table 1 discloses, that regression relationship performs quite respectably if asked to predict the volume of "full employment" GPS for the period after 1956 and for the years immediately antecedent to 1929

The ratios of the standard errors of the forecasts to the average levels of GPS for the forecast periods are shown in the addendum to table 1 The ratio for the forecast period 1957–64 studied by Hickman is about the same as the ratio of the standard error of the regression to mean GPS for 1948–56· the forecast ratio is 032, the estimation ratio is 034 The ratio of the standard deviation of the statistical discrepancy to mean GPS, however, is much smaller than the forecast ratio in this prediction period Because it is also smaller than its counterpart in the 1929/1948–56 period, the fall relative to the forecast ratio must be attributed primarily to the Commerce Department's increased accuracy rather than to any notable deterioration in the predictive performance of the Denison equation

The results for the extended forecast period 1957–69 suggest that there may have been some deterioration in the year-to-year stability of the GPSR from 1965 to 1969 For the extended period the forecast ratio is 054—over one percentage point greater than the estimation ratio for the 1929/1948–56 period This deterioration is the result of the historically high savings rates for 1966 and 1967, one has to go back to 1925 to find a GPSR which is higher Whether these sorts of extremes will recur remains an open question Certainly the longer-run stability of the GPSR seems unaffected by them, the average GPSR over the 1961–69 cycle is the same as the average for the 1954–57 cycle

The forecast ratios are reported for the period including the 1920s as well Only the "high employment" years for the twenties, that is, 1925, 1926, and 1928, are used Not surprisingly, adding only 3 years does not change things very much; the forecast ratio is raised from 054 to 058 For those years separately, however, the forecast ratio is 119, by comparison the ratio for the statistical discrepancy is 073 [3]

While there is evidence of greater year-to-year variability in the GPSR before the Second World War, there is no evidence of any longer-run instability In table 1 are reported the results of estimating the Denison relation for the entire period 1921–40/1948–64 and for the pre- and post-World War II subperiods A Chow test for the equality of coefficients between subperiods was performed, the test statistic was considerably below the 95 percent confidence limit, leading to the acceptance of the null hypothesis of no difference in the coefficients across subperiods If saving in favor of consumer durables, a conclusion in accord with the view that the 1920s marked a watershed in the widening of the range of consumer durables offered and their acceptance and consequent adoption by the public

---

[3] For the years prior to 1929 we have estimated GPS data from both the income side and the expenditure side using the work of Goldsmith (1953) and Kendrick (1961) The difference between the two sets was used to calculate the statistical discrepancy. A regression of one series on the other excluding the years 1917–20 yielded an insignificant intercept term and a slope coefficient of .98.

When restated in terms of plausible, more comprehensive savings and income variables, the "Denison relationship" can therefore be said to have persisted for an impressively long span of time. Indeed, there are indications that the long-run level of the GPSRA has remained unchanged during the past 100 years [5]

III

It is important to understand that the regularity in the GPSR which we have been at some pains to document in the last section is not the result of constant sectoral savings propensities, of a stable distribution of output between the public and private sectors, and, within the latter, of a stable division of income between the corporate and household sectors If the historical experience of the U.S. economy could be portrayed as a movement along a balanced growth path of that sort, our findings of a stable GPSR would be neither surprising nor very interesting

The facts, however, suggest a different picture The extreme stability of the GPSR (and the GPSRA) masks two rather important trends First, there has been a shift in the composition of private saving away from personal saving to corporate saving and expenditure on consumer durables Second, the total of private saving has been almost totally insensitive to the share of output absorbed by the public sector, given the level of output The first trend, in the context of a stable private saving rate, implies that wealth holders, for whatever reasons, behaved as if the components of saving were perfect substitutes The second also implies a high degree of substitutability, in this case between private consumption and taxes and between private investment and government dissaving Each of these results has implications which are far-reaching, intriguing, and important in terms of their application to public policy and economic theory

Cycle averages of the underlying components for the corporate and personal savings rates' contributions to the GPSR are listed in tables 3 and 4, respectively The trends in these components, summarized by pre-World War I, interwar, and post-World War II averages, reveal that there has been a redistribution of private saving away from personal saving narrowly defined and in favor of corporate saving and consumer durables expenditures. For example, the personal saving rate out of disposable personal income (DPI) fell by about 25 percent from the pre–World War I period to the period after World War I, while the corporate saving rate rose by over 20 percent. The fall in the personal rate is nearly two percentage points, over the same period the ratio of consumer durables expenditures to DPI increased by five percentage points.

---

[5] Although reliable annual estimates of output and saving are not available for the period before 1898, it is possible to construct decade averages of the GPSR and GPSRA for the period 1869–1908 using the work of Gallman and Davis (1968) and David and Abramovitz (1969). The estimates of the GPSR range from .173 to .191. For the GPSRA the range is from 216 to 233. The estimates for GPSR and GPSRA in the last decade, 1899–1908, are, respectively, 187 and .233. These compare with the estimated coefficients on GNP for the 1898–1916 period of .1835 and .2326, respectively.

The comparison of the personal saving rate and the rate of durables expenditures, however, is sensitive to the choice of base period. Thus, using the 1898–1916 period as a base suggests a shift in favor of durables, the fall in personal saving to DPI from this period to the post–World War II period is from .0804 to .0627, while CED/DPI (not shown) rises from .0657 to .0970. But the changes in the rates from the 1921–28 period to the present are both positive: a 1.37 percentage point change for the conventional savings rate and 2.19 points for the durables expenditures rate. Obviously, the important shift in the composition of personal saving (including consumer durables) comes at the end of World War I.[6]

One explanation for these large shifts is the response of savers to the differential tax treatment of income from different forms of investment. Thus, the secular rise in the corporate retention rate together with the decline in the narrowly defined personal saving rate suggest that the personal and corporate income tax schedules, along with the treatment of capital gains, the tax shelter features of depreciation provisions in the corporation income tax laws, and the like, have combined to effectively subsidize corporate saving at the relative expense of personal saving.[7]

While the high rate of substitution of corporate for personal saving shown in tables 3 and 4 may prompt some suspicion that the foregoing

---

[6] Thus, we ally ourselves with Oshima (1961) and Juster and Lipsey (1967) in finding a rise in share of durables spending in income after World War I. On the other hand, we find no evidence of a rising trend in the importance of durables thereafter, and in this respect we agree with Vatter and Thompson (1966) and Vatter (1967). In support of this last proposition, note that the augmented personal saving rate (PSR) is about half a percentage point higher in the post–World War II period than in the twenties and about a third of a percentage point higher than in the years before 1917. Neither of these differences is significant. Moreover, the PSR increased by a small amount between the twenties and the 1948–64 period, implying that the rise in the ratio of CED to GNP was very marginal. Indeed, across the three cycles following 1948, the decline of the augmented PSR, when the conventional rate fell by only half as much, supports Vatter and Thompson's point that the high spending on durables after World War II reflected backlogged replacement demand.

[7] On the other hand, the contrast between (1) the divergent behavior of the consumer durables expenditures rate and PSR out of DPI prior to 1929, when tax rates did not change very much, and (2) the parallel movement of these rates after 1929 as both corporate and personal tax rates rose, suggests that the shift in favor of durables during the 1920s was primarily a response to a change in the supply price of durables rather than to the effect of the heavier taxation of other income upon the relative after-tax rate of return on this class of investments.

TABLE 3

Components of Gross Corporate Saving Rate Out of GNP: Cycle Averages, 1898–1916/1921–29/1949–69

Cycle	Gross Corporate Saving to GNP	Gross Corporate Saving to Gross Corporate Profit after Tax	Gross Corporate Profit after Tax to Gross Corporate Profit before Tax	Gross Corporate Profit before Tax to GNP
1898–99*	1182	7863	1 0000	.1503
1900–3	1192	6932	1.0000	1725
1904–7	1041	6169	1 0000	1688
1908–10	1023	6101	0 9967	1681
1911–13	0964	5943	0 9940	1632
1914–16†	1221	6748	0 9888	1822
1921–23	1023	7107	0 9231	.1564
1924–26	1064	7407	0 9138	1573
1927–29	1047	6815	0 9183	1673
1949–53	1041	7947	0 6933	1875
1954–57	1123	8101	0 7325	1893
1958–60	1136	8126	0 7536	1857
1961–69	1147	8118	0 7608	1858
Period averages				
1898–1916	1104	6626	0 9966	1675
1921–29	1045	7110	0 9184	1603
1949–69	1112	8073	0 7351	1870

* Complete cycle 1897–99

† Complete cycle 1914–18

TABLE 4

Components of Personal Saving Rate Out of GNP Cycle Averages, 1898–1916/1921–29/1949–69

Cycle	Personal Saving to GNP	Personal Saving to Disposable Personal Income	Disposable Personal Income to Personal Income	Personal Income to GNP	Personal Saving Plus Consumer Expenditure on Durables to GNP Plus Gross Rental Flow on Durables
1898–99	0716	0877	9867	8249	1211
1900–1903	0771	0941	9863	8288	1283
1904–7	0783	0926	9894	8414	1330
1908–10	0685	0816	9885	8462	1212
1911–13	0528	0642	9886	8319	1115
1914–16	0522	.0624	9871	8429	1134
1921–23	0368	0449	9716	8399	1096
1924–26	0461	0571	9743	8293	1258
1927–29	0364	0451	9709	.8286	1169
1949–53	0476	0676	8915	.7928	1337
1954–57	0452	.0647	.8830	7917	1289
1958–60	0410	0583	8784	.7988	.1197
1961–69	0414	0601	8689	7925	1234
Period averages					
1898–1916	0668	0804	.9877	.8360	1214
1921–29	0398	0490	.9722	.8326	.1174
1949–69	0438	.0627	.8804	7940	.1264

cannot have been the sole influences at work, this is not nearly so strong an objection as it appears to be at first glance. When one considers the high degree of concentration of ownership in corporate debt and equity in the hands of those occupying the upper income brackets, the differences which tax laws tended to create in marginal after-tax rates of return could quite plausibly have provided the impetus for the marked portfolio shift observed.[8]

Whatever the reason for the shift in composition—and we return to the question of the role of taxes later—the intriguing point remains that the shift left the GPSR relatively undisturbed In other words, wealth holders behaved as if corporate and personal saving were near-perfect substitutes in the portfolio

Bailey (1972) argues that such behavior is consistent with households exhibiting a degree of ultrarationality in their dealings with the corporate sector which he calls "perfect foresight" Thus, households with perfect foresight perceive that a switch in corporate financing from retained earnings to debt issue, for example, does not affect their permanent incomes Their consumption therefore does not change, all that occurs is a switch from corporate to personal saving, with no change in the total

Certainly the behavior of corporate and personal saving is consistent with such a paradigm The question remains of whether it can be used to explain the two other examples of perfect substitution suggested by the data and mentioned earlier, namely, between consumption and taxes and between investment and the government deficit

These types of substitution are implied by the remarkable year-to-year stability in the GPSR despite swings in the amount of output absorbed by the government in taxes and debt issue, and the lack of a trend in the GPSR in the face of an inexorable rise in the size of the government sector. The stability of the GPSR despite changes in disposable income implies that consumption and taxes in effect are traded off one for one Its invariance to changes in the size of the government deficit suggests that private debt and public debt are close substitutes in private portfolios.

The application of Bailey's specification of ultrarationality to relations between the public and the government implies the same sort of

---

[8] Brittain's (1966) study of corporate dividend policy, for example, found that one of the major determinants of the desired payout rate (which the period averages in table 3 presumably come close to measuring) was the tax shelter offered to the highest decile of personal income tax payers by the combination of lower corporation income tax and capital gains tax rates Thus, a rise in the personal income tax rate for that class of taxpayers would lower the desired payout ratio, i e , increase corporate saving. The other major determinant of corporate saving that Brittain identified was the depreciation shelter provision of the corporation tax laws, thus, liberalized depreciation provisions tended to raise the retention rate out of net profits after tax. But this second influence may be allowed for in redefining the base—as Brittain and we have done—in terms of cash flow, i.e., gross profits after tax. It is also possible that the depreciation shelter provisions contributed to the rise in the ratio of gross corporate profits before tax to GNP which is observed after World War II

TABLE 5

STABILITY OF GROSS PRIVATE AND GROSS NATIONAL SAVINGS RATES, 1898–1971

	GROSS PRIVATE SAVING RATE			GROSS NATIONAL SAVING RATE		
PERIOD	Mean	Mean of Absolute Deviations of Annual Rates from Mean	Mean of Absolute Deviations as Percentage of Mean	Mean	Mean of Absolute Deviations of Annual Rates from Mean	Mean of Absolute Deviations as Percentage of Mean
1898–1928	16.77	2.79	16.64	16.92	2.82	16.67
1929/1948–64	15.46	0.41	2.65	15.51	1.29	8.31
1948–64	15.50	0.40	2.57	15.49	1.33	8.59
1948–71	15.62	0.48	3.07	15.46	1.17	7.57

Modigliani-Miller theorem applied to government finance that we saw for private finance. A switch from taxes to Treasury debt issue should leave consumption unaffected, while private saving varies directly, dollar for dollar, with the government deficit. On this construction of ultrarationality, private saving and public saving are perfect substitutes and national saving is the stable aggregate given the level of output. Yet that is at variance with the facts, which suggest that public saving and private investment are complements and that gross private saving is the more stable of the two aggregates in its relation to gross income.

We have adduced evidence that the GPSR is highly stable, both from year to year and over the long run. Some evidence on whether it has been more stable than the national saving rate is shown in table 5. Table 5 contains period means for both the national and the private savings rates as well as the mean absolute deviations of annual rates from their period means. For the period before 1929 the two rates are equally stable or unstable from year to year as measured by the ratio of their mean absolute deviations to their means. But in the period after World War II, the national rate is clearly the more unstable, its variability is from two and one-half to four times that of the GPSR. Over the long run the national rate appears marginally less steady, having been slightly larger than the GPSR before 1929 and somewhat smaller since then.

The notion of complete, ultrarational household portfolio management need not be abandoned, however, if a modification in Bailey's model is accepted. Suppose, to return to the analogy with pure private sector ultrarationality, that there is a shift in the type of firm expenditure coincident with the change in the method of financing. To be concrete, imagine that dividends are increased by reducing expenditure on non-monetary forms of executive compensation, such as free vacations, expense accounts, and the like, which are treated as consumption items by their recipients, while the new debt issue is used to finance capital expenditure. Hence, there is a shift in the time profile of income coincident with a change in the manner of financing. If households prefer the initial distribution of consumption, they will use the additional dividends to replace the consumption formerly provided by firms rather than spend them on the new debt issue. At the same time, because greater provision for future income is now being made in the corporate sector, households will reduce their own investment outlays. In other words, households absorb the new debt by reducing expenditure on internally financed investment; corporate investment and household investment are perfect substitutes in this example

Similarly, at the national level, if the private sector regards tax-financed government expenditures as belonging to the category of consumption, and bond-financed public outlays as investment, the marginal propensity to consume out of tax reductions is unity, and private saving rather than national saving is the stable aggregate. Bailey's dismissal of such behavior, which is exactly what we observe, as inconsistent with total rationality, stems from his implicit assumption that the uses to which public funds are put yield services that are perfect substitutes in private utility functions Given that assumption and perfect foresight, national saving logically is the stable aggregate But if consumption and investment (saving) are not perfect substitutes in household plans, and if the purposes of public expenditures are believed to be intimately tied to their mode of financing, then a stable private saving function is perfectly consistent with rationality.[9]

In recognition of imperfect knowledge, Bailey offers pragmatic criteria for the optimal financing mix for government expenditure The argument is roughly as follows Where the benefits and incidence of government expenditure are localized and certain, the cost of financing them should be distributed in the same way, which leads to a presumption favoring the tax financing of such types of expenditures By contrast, where the benefits are highly diffused or their incidence is uncertain, a bias is created in favor of debt financing, inasmuch as the costs of debt service similarly will be widely diffused and relatively uncertain If we interpret Bailey's dichotomy between certain and uncertain as a distinction between present and future consumption benefits, households conforming to our standard of ultrarationality will anticipate (or see to it) that their government in effect follows his recommendations: finance public investment with bonds, and pay for present benefits (consumption) with taxes.

---

[9] Bailey (1962, p 73), in an earlier discussion of ultrarationality, writes one version of the consumption function as

$$\frac{C + G_c}{P} = \frac{a}{P} + \frac{bY}{P}. \tag{1}$$

All items are in nominal terms, with P being the deflator, C the private consumption expenditure, G_c the consumption element in government expenditure, and Y the GNP If we identify G_c with T, total taxes, the GPS function implied by (1) is

$$S/P \equiv Y/P - (C + T)/P \equiv Y/P - (C + G_c)/P = -a/P + (1 - b)Y/P$$

Hence this variant of ultrarationality implies that GPS is a stable function of GNP and that at least the marginal GPSR is highly stable

This last prescription allows us to confront a problem which we have avoided up to now our version of ultrarationality is at odds with the explanation of the *shift in the* composition of private saving as an attempt to minimize the burden of taxes If households are ultrarational in the way we have described, and are indifferent as to whether consumption and investment are provided through the public or private sectors, why *should* the amount of taxes (and, for that matter, debt issue) be of any concern? Indeed, rather than indulge in tax avoidance, an ultrarational sector should act to equalize gross of tax rates of return

Tax avoidance can be reconciled with ultrarationality, however, if one assumes that the government does not follow the optimal finance mix policy and, instead, finances part of its investment program through taxes In such circumstances, tax avoidance can be interpreted as the private sector's attempt to shift the mix of financing toward the optimum directly, without engaging the cumbersome political machinery of the Congress Of course, if we are allowed thus to suppose that the political process is not wholly responsive, there is no assurance that legislative allocations will reflect private demands for public goods, the systematic avoidance of taxes could then be explained simply as the (private) attempt to rectify what is perceived as the overproduction of public consumption services Yet this is an answer too much in the spirit of Alexander's approach to knots We set out to see whether the phenomena of tax avoidance could be accounted for in a model which envisaged the corporate *and* public sector savings plans being integrated with and, at least for the long run, being under the control of an ultrarational household sector To depict the household sector as locked in conflict with an unresponsive government is to deny that the issue is worth bothering with, rather than to resolve it

Perhaps a more satisfactory, and certainly a more sophisticated, interpretation is one that recognizes the practical problems of matching the distribution of benefits of government expenditures with the incidence of the costs, and the consequent incidental redistributive effects of government fiscal operations Bailey's own discussion acknowledges that there is likely to be some conflict within the private sector between those who want publicly provided goods and those who must pay for them. And, as we have previously remarked, the appearance of (rational) integrated personal and corporate savings decisions designed to minimize the tax burden might well arise from the behavior of the wealthy few who effectively "own" the corporate sector. From the vantage point of these households, the amount of public goods provided may indeed be too large. The fact that the private sector is not composed of identical households is surely an important element of the story, but tax avoidance may be more than a symptom of nonoptimal, unintentionally redistributive modes of financing which concentrates the cost incidence of programs having widely diffused benefits It could also be a manifestation of the familiar "free rider," or "revelation of demand for public goods," problem. everyone may wish to see some publicly provided goods substituted for private consumer spending—and yet be reluctant to disclose their willingness to bear the necessary share of the incremental taxes.

IV

The particular conception of ultrarational household behavior that has been advanced here to account for the apparent absence of any significant long-run effects of tax changes on GPS, given the level of GNP, turns out to carry several rather strong fiscal policy implications These deserve to be spelled out in greater detail and contrasted with the weaker, but nonetheless significant, implications that flow more immediately from the existence of the Denisonian empirical regularities which we have sought to document and rationalize

Unlike Bailey's specification of perfectly foresighted behavior, our brand of ultrarationality implies that an extra dollar of government deficit will displace a dollar of private investment expenditure because households autonomously treat deficits as public investment and regard the public and private sectors' investment projects as interchangeable. We shall refer to this displacement effect as the "ex ante crowding out" of private investment [10] When output is at the full-employment ceiling, however, it is difficult to distinguish the foregoing phenomenon—at least as far as the impact on private investment is concerned—from what may be called "ex post crowding out" In the latter case, an output constraint, a stable demand function for money, no change in monetary policy, and a stable GPSR imply that an increase in the government deficit must displace an equal amount of private expenditure through a rise in interest rates—no matter how the households view public debt instruments and public investment projects

The difference between the two forms of crowding out is illustrated by the loanable-funds diagrams shown in figure 3*a* and *b* The vertical line SI_0 in each is the amount of GPS, and hence the supply of loanable funds, for an initial equilibrium level of output of Y_0 The fact of a stable GPSR is reflected in the complete interest inelasticity of saving.[11] Schedule II, the marginal efficiency of the investment curve, represents the demand for funds for investment Initially there is no government borrowing, and equilibrium occurs with the interest rate at r_0 and the level of

---

[10] Cf Friedman and Heller (1969, pp. 54–56) for a discussion of hypotheses about the "crowding out" effects of government bond issues.

[11] Wright (1969) presents evidence for a positive substitution interest elasticity of aggregate savings. But the estimates, while statistically significant, are not large; the highest estimate is 0.27. And Wright admits that the effect of interest rate changes on income makes the interest rate elasticity of the saving *rate* smaller. The high degree of stability of the GPSR that we have shown suggests that the interest elasticity of the savings rate is, for all practical purposes, zero.

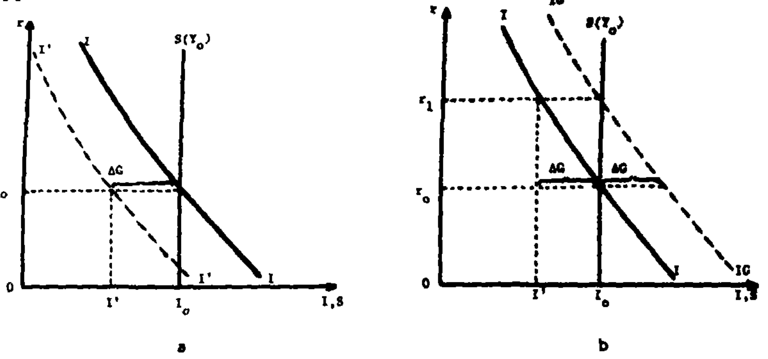


FIG 3

investment at OI_0 Now suppose the government borrows ΔG The ex ante case is shown in figure 3*a*, where the private sector treats the government deficit as part of investment and incorporates it into the investment schedule *II* The demand for private investment therefore declines by ΔG to $I'I'$ and the rate of interest is unchanged

The alternative form is illustrated by figure 3*b*, in which the addition of government borrowing of ΔG raises the total demand for loanable funds by that amount to *IGIG*. If there is some elasticity in the demand for or supply of money, the resulting rise in r generates an excess supply of real balances, part of which is attracted into the loanable-funds market The supply of loanable funds can thus differ in the short run from the funds supplied out of GPS, but in the full equilibrium where the demand and supply of the stock of money are perfectly matched, the supply of loanable funds it identical with GPS It is clear then, in terms of figure 3*b*, that the new equilibrium depends on what happens to SI_0 when the public sector goes into deficit

If Y_0 is less than full employment output, the increase in expenditure financed out of idle balances (or increases in the stock of money caused by a rise in the money multiplier) raises the level of aggregate output and hence shifts the SI_0 curve to the right The total of government and private borrowing is larger than before, and therefore the government deficit does not completely displace private investment.

Alternatively, if Y_0 is full-employment output, SI_0 is the fixed, full-employment level of saving An increase in government borrowing in this situation drives up the rate of interest to the point where private investment declines by the increase in the deficit. As is illustrated in figure 3*b*, in the final equilibrium the rate of interest has become r_1 and private investment has declined by $I'I_0 = \Delta G$. This is the case of ex post crowding out.

In terms of the impact on the level of investment, it is impossible to distinguish between the two types of crowding out. However, note that the ex post example assumed a given disposable income. The stable GPSR in that case can be rationalized in terms of a stable propensity to consume out of disposable income But the empirical evidence suggests that, in the case where the increase in the government deficit was at the expense of taxes, for example, the level of private saving still would not be affected In other words, the ex post crowding out can explain why private investment and government dissaving behave like perfect substitutes, but it is silent on why taxes and consumption should do so

One's view of the relative effectiveness of fiscal and monetary policy for macroeconomic stabilization, and the appropriateness of econometric exercises designed to test such views, turns out to depend in a fairly crucial way on whether crowding out can be taken to be of the ex ante, rather than the ex post, form Given that taxes displace private consumption expenditures, if private debt issues are displaced ex ante by government bonds, there is no way that changes in the level of public expenditures can affect aggregate demand and output (or prices), fiscal policy would be useless for stabilization purposes Nor would there be any additional "kick" of a larger multiplier attached to debt-financed government expenditure

To establish the presence of the ex ante form of crowding out, however, it is necessary to concentrate on the evidence provided by episodes in which there is sufficient slack in the economy to preclude the existence of ex post crowding out Even this empirical strategem is not totally dependable, for, if the money demand and money supply functions are extremely interest inelastic, considerable ex post crowding out may take place even in conditions of less than full employment Still, some econometric indications of the importance of monetary policy may be obtained by discarding observations relating to full-employment states, as in the latter it is difficult to distinguish between the effects of interest inelasticity of the demand and supply functions for money and the effects produced by a stable GPSR and the constraint on output

The absence of pronounced depressions, and the small year-to-year variation of the GPSR during the postwar era, does rather suggest that the disagreements which the monetarist-fiscalist debate has recently stirred up regarding the proper interpretation of the empirical record for the United States in this period may amount, after all, to little more than a squabble over residuals[12] The only way that controversy is likely to

[12] The article that started the whole debate is, of course, the now-classic piece by Friedman and Meiselman (1963). The evidence of a stable GPSR suggests, however, that their tests using consumption as the dependent variable are biased against the autonomous expenditure theory. The reason is that a stable GPSR implies that consumption *plus taxes* is the stable function of GNP, and hence of autonomous expenditure. Or to put it another way, the Friedman-Meiselman tests using consumption pick up not only any instability in the multiplier but the instability in the private consumption function as well.

be resolved is through investigations of the role played by monetary and fiscal policy in fairly severe disequilibrium states. This is a line of approach that has been pursued in recent time-series studies of monetary policy during the Great Depression (e g , Brunner and Meltzer 1968). But there is also room for international cross-section analysis of the workings of monetary and fiscal policy in less developed countries, where disequilibrium takes the form of being "unlike" the advanced industrialized economies The analogy drawn between the disequilibrium states of the industrially advanced countries and the prevailing situation in less developed countries is not so strained as it might first appear Leijonhufvud's (1968) work on the role of money in a general equilibrium context has suggested that a disequilibrium state can usefully be modeled as a situation in which information and transactions costs are at very high levels, and this seems a quite natural characterization of chronic (rather than transitory) conditions in the markets of many less developed countries

Of course, we are not confined to observing quantities alone, and the behavior of interest rates and prices when fiscal policy is altered—even in conditions of full employment—should provide information which permits ex ante crowding out to be distinguished from its ex post twin Thus, if a movement toward fiscal ease leaves prices and interest rates unaffected, this may be read as indicating that crowding out in the ex ante form had occurred Unfortunately, students of the postwar U S economy will find it difficult to isolate such episodes of "pure" fiscal policy changes The evident commitment of the Federal Reserve to the task of stabilizing interest rates "in support of" Treasury debt operations results in fiscal policy shifts being accompanied by alterations of the money supply But the latter's impact on prices and nominal interest rates (after these have adjusted to inflationary expectation) runs in the same direction as the price changes that would materialize in the course of ex post crowding out, thereby muddying an otherwise clear test

When one turns to the role of fiscal policy in the context of long-run growth, it may seem that the distinction between the two forms of crowding out with which we have been concerned is really unimportant In neoclassical growth theory, the supply constraint on output is assumed to be binding, in which case the constancy of the GPSR is enough to guarantee the same results with respect to the effect of government deficits or surpluses on private investment as ex ante crowding out. Yet it does matter When households act in accord with our definition of ultrarationality, the optimal mix of tax and debt financing for public expenditure—which the government either adheres to or finds that the private sector imposes by means of tax avoidance—introduces an additional constraint. This leaves fiscal policy no scope to affect the accumulation of capital, just as in the short-run the combination of ex ante crowding out and Denison's law robs fiscal policy of any leverage over the state of aggregate effective demand In the long run, fiscal measures can achieve an alteration of the distribution of activity between the private and public sectors, but, by hypothesis, that shift will not disturb the growth path of the economy It would seem that the stress that optimum growth theory (e g., Foley and Sidrauski 1971) has put on the role of fiscal policy as a control instrument is in need of some reconsideration When the economic relations between the household sector and its government are those which our standard of ultrarationality implies, optimal fiscal policy begins to appear just as neutral as money in long-run growth

The stability of the GPSR in the face of substantial financial innovation over the long span of time surveyed here might also suggest that the role of financial intermediation in American economic growth has been considerably less important than it has lately become fashionable to maintain But the temptation to jump to this revisionist conclusion should be resisted, for it unwarrantedly neglects the significant effects of financial intermediaries in facilitating the shifts in the composition of savings which, over the long run, have probably been essential in staving off a decline in the real rate of return on "capital "[13] The growth in mortgage financing of housing expenditures, for example, freed household internally generated funds for consumer durable purchases These funds in turn were supplanted by the growth of installment credit, which has allowed internal finance to be applied to other forms of capital accumulation, perhaps most notably now to education [14]

Whatever the reason for the crowding out of private investment—"lack of perfect foresight" as Bailey would have it, "ultrarationality" as we have defined that, or just the conjunction of full employment and the stable GPSR required by Denison's law—the empirical findings reviewed here offer a simple and direct answer to the important public finance question What rate of discount should be applied in evaluating public investment projects and, more generally, to government borrowing for any purposes? The historical evidence of a highly stable GPSR suggests that the interest elasticity of saving is very low and may be approximated by zero That being the case, as Harberger (1971) has shown and as one can plainly see from figure 3, we can say that the opportunity cost of the resources borrowed by the government (ΔG) is unambiguously represented by the area under the marginal efficiency of investment schedule II in the interval ΔG between I' and I_0. Since the appropriate discount rate is given therefore by the social marginal productivity of investment in the private sector, it would seem that the awkward problems of measuring the social rate of time preference can be justifiably finessed Indeed, if households really were ultrarational in the way we have described, the marginal rate of time preference would be equated to the gross-of-tax social rate of return on private sector investments

---

[13] Cf Abramovitz and David (1973*a*, 1973*b*) on the capital-deepening bias of technological change and the changing composition of capital formation over the course of the past and present century of growth in the United States

[14] In his study of the demand for money Chow (1966) found that permanent income, which presumably includes the return to human capital, was a better explanatory variable than nonhuman wealth alone However, the relationship between financial intermediation and the growth of human capital mentioned above raises the question of whether something other than a demand relationship is being measured. For the narrow definition, i.e., including only commercial bank demand deposits, this may not be important But it is not clear that studies of the demand for more inclusive definitions of money are not contaminated by some simultaneity bias of this sort

References

Abramovitz, Moses, and David, Paul A "Economic Growth in America Historical Parables and Realities " *De Econ* 121 (May/June 1973) 251–72 (*a*)

——— "Reinterpreting Economic Growth Parables and Realities of the American Experience " *A E R* 63 (May 1973) 428–39 (*b*)

Ando, Albert, and Modigliani, Franco "The 'Life Cycle' Hypothesis of Saving Aggregate Implications and Tests " *A E R* 53 (March 1963) 55–84

Arak, M V , and Spiro, Alan "The Relationship between Permanent Income and Measured Variables " *J P E* 79 (May/June 1971) 652–60

Bailey, Martin J *National Income and the Price Level* New York McGraw-Hill, 1962

——— "The Optimal Full Employment Surplus " *J P E* 80 (July/August 1972) 649–61

Bhatia, Kul B "Accrued Capital Gains, Personal Income and Saving in the United States, 1948–1964 " *Rev. Income and Wealth,* ser 16, no 4 (December 1971), pp 363–78

Brittain, John A *Corporate Dividend Policy* Washington Brookings Inst , 1966

Brunner, Karl, and Meltzer, Allan "Liquidity Traps for Money, Bank Credit and Interest Rates " *J P E* 76 (January/February 1968) 1–37

Chow, Gregory C "On the Long-Run and Short-Run Demand for Money " *J P E* 74 (April 1966) 111–31

David, Paul A , and Abramovitz, Moses "Economic Growth in the United States Savings in the Nineteenth Century " Unpublished Stanford Center Res Econ Growth, August 1969

Denison, Edward F "A Note on Private Saving " *Rev Econ and Statis* 40 (August 1958) 261–67

Duesenberry, James S *Income, Savings, and the Theory of Consumer Behavior* Cambridge, Mass Harvard Univ. Press, 1949

Foley, Duncan, and Sidrauski, Miguel *Monetary and Fiscal Policy in a Growing Economy* New York Macmillan, 1971

Friedman, Milton *A Theory of Consumption Function* Princeton, N J. Princeton Univ. Press, 1957

Friedman, Milton, and Heller, Walter W. *Monetary vs Fiscal Policy.* New York Norton, 1969

Friedman, Milton, and Meiselman, David. "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897–1958." In *Stabilization Policies* New York: Prentice-Hall, 1963

Friend, Irwin, and Schor, Stanley. "Who Saves?" *Rev. Econ. and Statis.* 41 (May 1959): 213–48.

Gallman, Robert E., and Davis, Lance E "The Share of Savings and Investment in Gross National Product during the 19th Century, United States of America " Memo no 63. Unpublished. Stanford Center Res. Econ. Growth, July 1968.

Goldsmith, Raymond W. *A Study of Saving in the United States.* Princeton, N.J : Princeton Univ. Press, 1953.

Grunfeld, Yehuda, and Griliches, Zvi. "Is Aggregation Necessarily Bad?" *Rev Econ and Statis.* 42 (February 1960) 1–13.

Harberger, Arnold C. "On Measuring the Social Opportunity Cost of Public Funds " *Inst Defense Analysis Econ. Papers* (May 1971), pp. 1–45

Hickman, Bert G "Investment Demand in the Sixties " Reprint no 120. Washington Brookings Inst , 1966

Juster, F Thomas *Household Capital Formation and Financing, 1897–1962.* Princeton, N J Princeton Univ Press (for Nat Bur Econ Res), 1966

Juster, F Thomas, and Lipsey, Robert E "A Note on Consumer Asset Formation in the United States " *Econ J* 77 (December 1967) 834–47

Kendrick, John W *Productivity Trends in the U.S* Princeton, N J Princeton Univ Press (for Nat Bur Econ Res), 1961

Leijonhufvud, Axel W B *On Keynesian Economics and the Economics of Keynes* New York Oxford Univ Press, 1968

Modigliani, Franco, and Brumberg, Richard "Utility Analysis and the Consumption Function An Interpretation of Cross-Section Data " In *Post-Keynesian Economics*, edited by K K. Kurihara New Brunswick, N J Rutgers Univ Press, 1954

Oshima, Harry "Consumer Asset Formation and the Future of Capitalism." *Econ J* 71 (March 1961) 20–35

Vatter, Harold, G "Consumer Asset Formation in the United States A Rejoinder," *Econ J.* 77 (December 1967) 848–55

Vatter, Harold G , and Thompson, Robert L. "Consumer Asset Formation and Economic Growth The United States Case " *Econ J* 76 (June 1966) 312–27

Wright, Colin "Saving and the Rate of Interest " In *The Taxation of Income from Capital*, edited by Arnold C Harberger and Martin J Bailey Washington Brookings Inst., 1969

The Cost of School Time, Foregone Earnings, and Human Capital Formation

Donald O. Parsons

Ohio State University

A simple educational investment model is used to demonstrate that, if students are subject to borrowing constraints, foregone earnings are not identical to schooling time costs, since students will sacrifice leisure as well as earnings. Direct measurement of schooling hours and work hours of young males reveals that at the high school level the bulk of school hours result from foregone leisure. A review of the foregone-earnings measures used in a number of major human capital studies is undertaken to determine the approximate bias in human capital formation and rate-of-return estimates resulting from this source.

This paper deals with theoretical and empirical aspects of the proper measurement of school time costs. Using foregone-earnings measures, most researchers have concluded that schooling time costs are the bulk of all school-related human capital investments. Schultz (1960, p. 580) estimated such costs as about 60 percent of all costs in 1956, while Becker's (1964, p. 75) estimate is 74 percent for private college education costs. The major point developed below is that indirect measures of the cost of school time such as "foregone earnings" seriously underestimate school time costs.

As a practical matter, there are three major components of schooling costs: direct expenditures on tuition, books, transportation, etc., the alternative value of the hours devoted to the schooling activity, and the reduced earnings on any hours the student continues to work due to part-time work wage penalties. In most of what follows, the emphasis will be on the estimation of schooling hours, although the size of the part-time work differential will also be estimated.

This research was supported in part by the Manpower Administration, U.S. Department of Labor, and the Ohio State University Computer Center. The comments and criticisms of Belton Fleisher, T. W. Schultz, and an anonymous referee are gratefully acknowledged. Research assistance was provided by Donald Smith, Jill Chappell, and Howard Atkins.

Virtually all major human capital studies have used earnings differentials of youth out of and in school as an indirect measure of school time costs, since earnings data are considerably more accessible than direct time allocation data [1] Implicit in this indirect measure of schooling time costs is the proposition that the difference in hours worked between the two groups is equal to schooling hours or, from the individual's viewpoint, that the student will reduce his work hours by a quantity identical to his schooling hours

This work/schooling assumption is untenable, however, as it appears that students as a group devote substantially more time to market activities (work plus schooling) than do nonstudents of the same age and level of schooling completed To the extent that students are generally more capable at market activities, of course, one might expect them to devote more time to market activities throughout their lifetimes This tendency is not particularly strong at later points in the life cycle,[2] but other factors may augment the effect for students in school As will be developed at greater length in Section I, borrowing difficulties during the educational investment period induce students to sacrifice leisure as well as work hours to attend school, since consumption cannot be financed by future earnings

In Section II, some evidence is presented on the extent to which schooling hours are drawn from work hours and from leisure hours For high-school-age youth, for whom the direct schooling hours data are richest, evidence indicates that substantial work time and leisure time are both sacrificed for school hours on average Evidence on the wage rate differential of youth in and out of school is also presented The principal data source is the National Longitudinal Surveys, a 5-year panel survey of (among others) about 5,000 male youth 14–24 years of age in the first survey

A comparison of methods of calculating school time costs used in several important human capital studies is presented in Section III Data problems have not even allowed the uniform measurement of the indirect

---

[1] See Section III below for a more detailed description of measurement methods in a number of major human capital studies

[2] Becker (1964, p 169) was aware of the problem that work hour differences between schooling groups may affect the interpretation of differences in earning streams and rate-of-return estimates He presented evidence which suggested that among employed workers there is no strong relationship between hours of work and education level Lindsay (1971) considers this same issue at greater length A regression run on data obtained from the National Longitudinal Surveys found the following for men age 45–59

$$\text{work hours} = 2{,}022 + \underset{(4\ 6)}{13\ 7}\ \text{years of school},\ \bar{R}^2 = 015.$$

This suggests that a worker who has completed 4 years of college will work on average 55 hours a year, or something less than 3 percent, more than one who has completed only high school.

measure of school time, foregone earnings, so the actual methods and assumptions used are applied to a uniform data base to determine the degree of bias of each from directly measured school hours With some exceptions, the foregone-earnings technique led to significant underestimation of the investment of time in schooling activities of males, particularly at the high school level The study which best estimates foregone earnings most seriously underestimates school time costs

I. A Theoretical Model of the Schooling-Work-Leisure Decision

Schooling choice models in a perfect capital market are well established in the human capital literature Ben-Porath (1967) presented an elegant multiperiod schooling-work model, while Becker (1967*a*) developed a similar model incorporating leisure choice as well If one assumes that an individual in school will undertake both leisure and work in that particular period and that assorted interperiod interactions are not large, there is some justification for assuming that schooling activities will substitute only for work hours, as the foregone-earnings approach requires Under such a set of assumptions, the constant marginal value of work hours, namely, the wage rate, makes work hours a buffer, with the value of school hours and leisure hours being marginally equated to the wage rate in equilibrium

A shift up in the schooling returns schedule in a given period will disturb the equality of the wage rate and discounted marginal returns to schooling, inducing more schooling Ignoring indirect effects, the equality of the constant wage rate and value of leisure time will be unaffected, suggesting that leisure hours will remain unchanged and work hours only will be reduced Obviously a shift up in schooling productivity is likely to change long-run wealth and future-period wage rates which are themselves likely to affect demand for leisure, but these may be second-order effects

One difficulty recognized by both Ben-Porath and Becker is that there will be a tendency to acquire education early in the life cycle, since the finiteness of life ensures that the stream of returns to a given educational investment will be larger the earlier the investment is undertaken In fact, without the assumption of sufficient diminishing schooling productivity within periods as well as cumulatively, a complete specialization of the individual in schooling and leisure would be the typical case As Becker (1967*a*, p. 14) has noted in such cases, the wage rate is no longer a useful index of the value of time[3] and, moreover, leisure hours would be strictly and negatively dependent on school time productivity.

---

[3] This is rather obvious, since the individual finds his working wage sufficiently unattractive that he chooses not to work. Gronau (1971) considers this question at great length in the context of valuing the time of housewives.

Individuals who find schooling more productive will tend to reduce their consumption of leisure, *ceteris paribus* Even in traditional educational investment models, then, foregone earnings will underestimate schooling time costs if a youth does not work part-time

Students, however, do typically undertake part-time work A sample of about 5,000 males age 15 to 25 in 1967 reveals that the age-work hours profile for full-time students (where full-time student status was self-declared) is approximately linear, annual work hours rising from under 500 hours at age 15 to almost 900 hours at age 23 (see table 1 for details) As mentioned above, Ben-Porath and Becker suggest that the reason for part-time work by students is that the efficiency of schooling investments decreases within time intervals as well as cumulatively across periods Since increased work hours are also likely to affect adversely the efficiency of a fixed number of school hours, the predicted substitution of work hours for school hours in this situation is not obvious An alternative explanation of the existence of part-time work among students is that students frequently face rather severe capital constraints [4] The assumption that students can borrow and lend at some set interest rate is intuitively unappealing It seems more plausible to assume that individuals have great difficulty in borrowing to finance human capital investments and instead are forced to self-finance their schooling by working part-time [5] Below I construct a schooling model to take account of this borrowing difficulty in a simple way and examine the effects of this modification on the schooling, work, and leisure choices

To formalize the discussion, consider an n-period model in which lifetime utility is a function of the stream of leisure hours and consumption over the n periods, so

$$U = U(L_1, L_2, \quad , L_n, X_1, X_2, \quad , X_n), \tag{1}$$

where L_i is leisure hours in the ith period and X_i is consumption of a composite market good in the ith period Assume that past schooling investments increase the ith period wage rate (w_i) in the following way

$$w_i = f(S_i)w_0, \tag{2}$$

where

$$S_i = \sum_{j=1}^{i-1} s_j,$$

---

[4] Becker (1967*b*), Stafford (1968), and Gustman (1969) consider various aspects of the borrowing difficulties of youth The complex problems of family support levels are considered in a separate paper

[5] Empirical tests of these two different hypotheses are not hard to construct The capital constraint model would suggest, for instance, that summer school would be attended by students from generally wealthier families. The secular increase in average daily attendance of school youth in the United States, a rise from 99 to 159 days between 1900 and 1957 for students 5–15 (see Schultz 1962, p. 97), could be explained by the general rise in affluence under the capital constraint hypothesis, while the intraperiod efficiency hypothesis would presumably require some technology argument.

with s_j the time units of schooling undertaken in the jth period and w_0 the initial-period wage rate. Presumably $f' > 0$ and $f'' < 0$

For simplicity, consider the polar case where no borrowing is assumed, all consumption and other expenditures undertaken by students must be financed by past or present earnings or assets In its most general form, the capital constraint can be represented by a set of n inequalities that ensure that the individual will never be a net borrower, namely, the following.

$$A_0 + \sum_{i=1}^{t} \delta^{i-1} \left[w_i(H - L_i - s_i) - PX_i - cs_i\right] \geq 0, \qquad (3)$$

$$\text{for } t = 1, 2, \quad , n,$$

where $\delta = 1/(1 + r)$ and r should be interpreted strictly as the lending rate, A_0 is the individual's initial asset level, H is the total hours available per period, c is the level of direct schooling costs (assumed proportional to school hours), and P is the price level of market goods

The objective of the individual is to maximize expression (1), his utility function, subject to the capital constraints in the n periods represented by the inequalities in (3) The Lagrangian appropriate to this case is

$$J = U(L_1, L_2, \quad , L_n, X_1, X_2, \quad , X_n)$$
$$+ \sum_{t=1}^{n} \lambda_t \left\{A_0 + \sum_{i=1}^{t} \delta^{i-1} \left[w_i(H - L_i - s_i) - PX_i - cs_i\right] - A_t\right\}, \qquad (4)$$

where the λ's are Lagrangian multipliers and the A_t's are nonnegative slack variables introduced to ensure equality of inequality set (3) Intuitively, λ_t represents the marginal utility of a given wealth level in a particular period In a model with borrowing and lending at a fixed rate, only λ_n would be greater than zero, since lifetime wealth and not the particular time stream of wealth position would be relevant The quantity A_t represents net nonhuman assets at the end of the tth period and is constrained in this problem to be nonnegative because borrowing is not allowed

When both leisure and market goods are consumed in all periods, as would typically be the case, the Kuhn-Tucker conditions for a maximum utility level are the following set of equalities and inequalities for each period.

$$\frac{\partial J}{\partial X_t} = U_{x_t} - \sum_{i=t}^{n} \lambda_i \delta^{t-1} P = 0, \qquad (5)$$

$$\frac{\partial J}{\partial L_t} = U_{L_t} - \sum_{i=t}^{n} \lambda_i \delta^{t-1} w_t = 0, \qquad (6)$$

$$\frac{\partial J}{\partial s_t} = \sum_{i=t+1}^{n} \lambda_i \left[\sum_{j=t+1}^{i} \delta^{j-1} \frac{\partial w_j}{\partial s_t} (H - L_j - s_j) \right] - \sum_{i=t}^{n} \lambda_i \delta^{t-1} (w_t + c) \leq 0, \tag{7}$$

$$\frac{\partial J}{\partial \lambda_t} = A_0 + \sum_{i=1}^{t} \delta^{i-1} [w_i(H - L_j - s_j) - PX_i - cs_i] - A_t = 0, \tag{8}$$

$$\frac{\partial J}{\partial A_t} = -\lambda_t \leq 0, \tag{9}$$

where the inequalities in (7) and (9) hold only if the related control variable is optimally zero. The expression (9) suggests that if the optimal strategy has the individual end the period with positive physical assets, $\lambda_t = 0$ or equivalently the constraint in that period can be ignored.

Interpreting the results, equations (5) and (6) suggest that an individual making a purchase or consuming leisure (thereby foregoing income) must weight the cost of the activity by the reduced options in all future periods in which the capital constraint is effective. Consumption today has implications for consumption in specific later periods as well as over the lifetime.

The schooling expression is simply a modification of the traditional rule that schooling should not be undertaken whenever the net discounted flow of incremental income is less than the sum of school-related costs. The only difference is in the definition of discounting, since borrowing is not possible. In periods where borrowing would have been undertaken, the individual is forced to self-finance expenditures and the value of earnings exceeds the lending discount factor. Rearranging expression (7) in terms of cash flows by periods, one has

$$-\left(\sum_{i=t}^{n} \lambda_i\right) \delta^{t-1} (w_t + c) + \left(\sum_{i=t+1}^{n} \lambda_i\right) \delta^{t} \frac{\partial w_{t+1}}{\partial s_t} (H - L_{t+1} - s_{t+1}) + \quad + \lambda_n \delta^{n-1} \frac{\partial w_n}{\partial s_t} (H - L_n - s_n) \leq 0. \tag{10}$$

Occurring first, schooling costs will always be weighted more heavily than future returns if capital difficulties exist, leading to the intuitively obvious result that borrowing difficulties reduce the profitability of schooling in a given period. An individual may "drop out" of school to earn money sufficient to eliminate the capital constraint but will still generally find less schooling profitable than his counterpart who, because of initial wealth or borrowing possibilities, invested in schooling earlier in the life cycle.

The main justification for developing the capital constraint model here is to acquire some notion of the substitutability of leisure and schooling hours. The model suggests that schooling time will come from leisure activities as well as work hours, whether work hours which would be undertaken are estimated from the behavior of comparable youth out of school or from the same youth after schooling is completed. That leisure time in equilibrium will be more valuable to a youth in school than a comparable youth out of school (assumed to be subject only to the lifetime wealth constraint) can be seen by the ratio of marginal utilities of leisure of students to those of nonstudents

$$\frac{U_{L_t}^{\text{student}}}{U_{L_t}^{\text{nonstudent}}} = \frac{\sum_{i=t}^{n} \lambda_i}{\lambda_n} = 1 + \sum_{i=t}^{n-1} \frac{\lambda_i}{\lambda_n} > 1 \qquad (11)$$

if at least some of the λ_i are positive Obviously, the full wealth positions of the student and nonstudent are not likely to be identical, so relationship (11) is, at best, a tendency, but one which raises some doubts about the validity of any simple application of the foregone-earnings concept

Longitudinal data which would allow one to obtain data on work behavior during and after schooling for the same individual would not necessarily lead to a more satisfactory measure of schooling hours It is easy to demonstrate that, if one ignores the complicating factor of direct school costs,

$$\frac{U_{L_t}}{U_{L_{t+k}}} > 1$$

if a capital constraint is effective in period t In fact, the time path of the marginal utility of leisure will fall over the lifetime in this simple model, suggesting that, if the taste for leisure does not vary between periods, leisure will increase over the life cycle Clearly, non-investment-related variations in wage rates over the life cycle, due, say, to maturation or physical depreciation, will modify this simple time structure The possibility remains, however, that schooling hours may come from leisure hours as well as work hours, suggesting that foregone earnings will systematically underestimate school time costs.

II. The Value of School Time: Some Empirical Evidence

The schooling model developed in the last section casts some doubt on the assumption that lies behind the use of foregone earnings as a proxy for value of school time, namely, that an additional hour of school implies one less hour of work Direct estimates of schooling time and work time are developed below which will allow estimation of the percentage of schooling time that is the result of foregone leisure. Wage rate functions are also estimated separately for youth in school and out to measure the wage rate penalty that students, typically part-time workers, must bear. This is a second dimension of schooling activity costs.

Foregone Work and Leisure Hours

Both intuition and the work hours data in the sample of male youth examined here indicate that youth in school work substantially fewer hours than do otherwise comparable youth out of school. Annual work hour estimates for male youth are presented in table 1 for out-of-school youth, full-time students, and part-time students. The work hour–age profiles were estimated as functions quadratic in age. Work hours are presented as an age profile because of the strong dependence of age on work hours, presumably resulting from a combination of family support trends, changing social pressures with age, and growing levels of maturity. The interesting question, of course, is the nature of the trade-off between work and schooling, not simply whether they are interdependent. To what extent are school hours derived from leisure hours?

To answer that question, a measure of schooling hours is needed. Such information is more readily available for high school activity than for college or university study. Schooling hours are composed of two major components, classroom hours and homework time. The Coleman report (Coleman 1966) presents data obtained from principals on days in session, length of academic day, and attendance rates for high school seniors (and certain lower classes as well) from which one can derive a measure of annual hours spent in class per student. The principals reported the figures presented in table 2 for the schools attended by white children.[6] The deviations from the mean are proportionately quite small for all but daily hours. Average annual hours in class can be estimated then as

$$\begin{aligned}\text{annual school class hours} &= \text{annual days} \times \text{daily attendance rate} \\ &\quad \times \text{hours per day} \\ &= (179.4)(0.943)(5.8) = 981.1,\end{aligned}$$

or approximately 1,000 hours.

The National Longitudinal Survey asks a question of all high school students and youth immediately out of high school on hours spent per week on homework, the second major component of schooling hours. The results are tabulated below in table 3. Assuming mean values for each interval and a mean of 22 for the open-ended class, mean weekly homework hours of 7.9 can be computed. For the approximately 40 school weeks per year, these figures would suggest annual homework hours of 316. Combining class hours of 981 with estimated homework hours of 316, we can derive an estimate of 1,297, or approximately 1,300 hours

---

[6] These figures are largely confirmed by other sources. The average legal minimum hours of school for those states (34) which have such laws are 5.86 per day in 1970. In 23 states, the legally required days per year are precisely 180 days (see *Educational Commission of the States* 1970). Average daily membership and average daily attendance for all public elementary and secondary schools are also available (National Education Association 1970). From these, the rate of absence can be calculated for the same period as 6.4 percent.

TABLE 1
ANNUAL WORK HOURS OF MALE YOUTH, 1967*

Age	In School Full-Time (1)	In School Part-Time (2)	Out of School (3)
14	282	3	494
15	402	366	768
16	508	694	1,018
17	599	986	1,246
18	677	1,242	1,450
19	741	1,463	1,631
20	791	1,648	1,789
21	827	1,797	1,923
22	850	1,911	2,035
23	858	1,990	2,123
24	852	2,032	2,187

* Data are from the National Longitudinal Surveys. Annual work hours were estimated from the following regressions

$$\text{work hours} = \underset{(4.59)}{-2{,}863} + \underset{(4.74)}{322.4}\ \text{age} - \underset{(3.83)}{6.98}\ \text{age}^2,\ R^2 = .069, \tag{1}$$

$$\text{work hours} = \underset{(3.19)}{-8{,}822} + \underset{(3.18)}{879.7}\ \text{age} - \underset{(2.64)}{17.81}\ \text{age}^2,\ R^2 = .390, \tag{2}$$

$$\text{work hours} = \underset{(5.77)}{-5{,}785} + \underset{(6.27)}{611.3}\ \text{age} - \underset{(4.97)}{11.63}\ \text{age}^2,\ R^2 = .197, \tag{3}$$

where absolute values of *t*-ratios are in parentheses.

TABLE 2
FIGURES USED IN CALCULATION OF AVERAGE CLASS HOURS BY HIGH SCHOOL STUDENTS*

	Mean	Standard Deviation
Annual days in session	179.4	4.03
Average daily attendance (%)	94.3	2.38
Length of academic day (in hours)	5.8	1.12

* Relevant to white students. Reported by principals (Coleman 1966, 2.73).

TABLE 3
WEEKLY HOMEWORK HOURS IN HIGH SCHOOL, MALE YOUTH, 1966

	HOURS					
	0	1–4	5–9	10–14	15–19	20 and More
Frequency (%)	2.9	26.7	36.9	23.2	7.2	2.9

SOURCE.—National Longitudinal Surveys.

for an estimate of total schooling hours for high school youth. For lack of data this figure will also be used as an estimate of college hours, although any conclusions drawn for college age youth must be considered extremely tentative.[7]

---

[7] One single college survey of homework hours suggested that college students spend considerably fewer hours at schooling activities than do high school youth (see Marwardt and Sikkink 1970). Measurement of the investment component of social interaction (bull sessions, etc.) would be necessary before any firm conclusions could be drawn, however, on this issue.

TABLE 4
SCHOOLING HOURS AND WORK HOURS

Age	Average School Hours (1)	Foregone Work Hours* (2)	Work Hours Share† (3)
14	1,300	212	16 3
15	1,300	366	28 2
16	1,300	511	39 3
17	1,300	647	49 7
18	1,300	773	59 4
19	1,300	890	68 4
20	1,300	998	76 7
21	1,300	1,096	84 3
22	1,300	1,185	91 2

* From table 1, column 3 minus column 1
† Column 2 divided by column 1 times 100

It is now possible to estimate what share of schooling hours comes from reductions in work hours In table 4, column 1, estimated schooling hours are reported while in column 2 the difference between work hours of nonstudents and work hours of students or "foregone work hours" is computed Column 3 then provides an estimate of the percentage of total school hours which are foregone work hours Quite clearly, for the younger age levels (where the schooling time estimates are most relevant) the bulk of schooling hours are due to sacrificed leisure rather than sacrificed work hours, or ,conversely, the bulk of school time released when a young male drops out of high school is directed (voluntarily or otherwise) into leisure time If wage rates for youth in school and out-of-school youth were comparable and the most simplistic measure of foregone earnings were used as the measure of schooling hours, these estimates would be radically wrong Wage rates, however, are not equal for youth in and out of school (even with equal levels of schooling and age), and, moreover, because of data limitations on earnings, most foregone-earnings measures are not so straightforward Foregone-earnings measures used in a number of important human capital studies are considered in detail in Section III to determine the likely bias they contain It turns out that in most cases the bias is considerably reduced because of inaccurate assumptions about the magnitude of student earnings Before considering this matter, I consider first the nature of the wage rate process

Unit Value of Time[8]

Allocating hours from work to school activity creates a secondary problem in that students face considerably lower wage rates for their typically part-time work than they would face if they were seeking full-time

---

[8] For a more detailed analysis of wage rate functions for a subset of the individuals considered here, see Parsons (1972).

TABLE 5
HOURLY WAGE RATE REGRESSIONS OF MALE YOUTH, 15–25 IN 1967*

Group	Constant	Log School Years	Log Age	R^2	Sample Size
Nonstudents	−3 692 (9.39)	0 678 (8.96)	0 920 (6 91)	186	768
Students	−2 658 (12 23)	1.205 (13 31)		116	1,341

SOURCE —National Longitudinal Surveys
* Dependent variable is the logarithm of hourly wage rates Calculated hourly wage rates greater than $7 50 are deleted The absolute value of the *t*-ratios are in parentheses

TABLE 6
HOURLY WAGE RATE POSSIBILITIES IN AND OUT OF SCHOOL FOR MALE YOUTH, 1967*

Schooling Level (Age)	Wage Rate In School	Wage Rate Out of School
9 (14)	$0 99	$1 25
10 (15)	1 12	1 43
11 (16)	1 26	1 62
12 (17)	1 40	1 82
13 (18)	1 54	2 03
14 (19)	1 69	2 24
15 (20)	1 83	2 46
16 (21)	1 98	2 69
17 (22)	2 13	2 92

* Calculated from estimates in table 5

employment This results presumably from the lower productivity part-time workers have to firms as well as the reduced incentive the part-time worker has to search out alternative employers and higher wages The precise size of this wage rate differential is a matter of some interest, since it will affect direct measures of foregone earnings. Part of the earnings foregone will be due to reduced wage rates of students rather than reduced hours.

Wage rates as a logarithmic function of years in school and age are presented in table 5 for male nonstudents and students. The age variable was not introduced in the student wage rate regression, since school year and age are extremely highly correlated for youth still in school Assuming the youth followed the normal course of completing each grade in 1 year, so age = schooling years + 5, one can use the information in table 5 to estimate the youth's wage rate options in and out of school at various points in the schooling process. The average wage rate options in and out of school at various years of schooling are given in table 6. In percentage terms, wage rates of in-school youth vary from about 79 percent of what they could have earned had they decided to work full-time at age 14 to about 73 percent at age 22. This penalty forms another incentive to specialize in schooling and allocate work time to the postschooling period.

III. Estimates of Foregone Earnings in Selected Human Capital Studies

A number of important human capital studies have estimated foregone earnings at different levels of schooling Because of data problems, however, most have not derived their estimates from direct measurement of differences in work hours between nonstudents and students The closest any study has come to that is the corresponding measurement of differences in earnings, presented in a recent paper by Solmon (1971) Ironically, Solmon's estimates, while more precise measures of earnings differences, are more seriously biased estimates of value of school time than any of the prior studies operating with more limited data The main purpose of this section is to compare methods of estimating the value of school time used in a number of human capital studies with a single data set This should give us an idea of the accuracy of the methods used relative to the directly measured 1,300 hours of school

Consider the following alternative methods of estimating the value of school time

Method 1 The Actual Difference in Earnings of Students and Nonstudents

As mentioned above this approach is used by Solmon (1971) and is purported to be a refinement of early estimates by Schultz (1960) Restricted to the hours allocation aspect, this approach involves computing the difference in work hours between students and nonstudents Estimates of this measure appear in table 4, column 2

Method 2 Three-fourths Earnings of Nonstudents

This measure was used in the classic rate-of-return studies by Becker (1964) and Hanoch (1967), where the particular fraction was derived from two other studies of student-nonstudent earnings differentials (see Becker 1964, p 170) Although really only another way of measuring the same quantity measured more directly by method 1, practically it yields different results, since full-time students typically work more than one-fourth as much as nonstudents

Method 2a The Full Earnings of Nonstudents

Justified purely on the grounds of simplicity, this measure is used by Hansen (1963) in deriving the value of school time.[9] If students should

---

[9] To quote Hansen (1963, p 130) "For example, at age eighteen the opportunity cost for the person undertaking four years of college is the income that the high school graduate would obtain from ages eighteen to twenty-one This procedure made it unnecessary to rely upon indirectly estimated opportunity cost figures . ."

TABLE 7
WEEKLY WORK HOURS, MALE YOUTH, 1966*

Age	All Youth (1)	Nonstudent Youth (2)
14	19.6	31.8
15	24.2	33.7
16	28.4	35.5
17	32.1	37.2
18	35.3	38.7
19	38.0	40.1
20	40.2	41.3
21	41.9	42.4
22	43.1	43.4

* Data are from the National Longitudinal Surveys. The age profiles were derived from the following estimated relationships

$$\text{hours per week} = \underset{(11.1)}{-97.6} + \underset{(12.5)}{11.83}\,\text{age} - \underset{(10.0)}{0.25}\,\text{age}^2,\ R^2 = .237, \qquad (1)$$

$$\text{hours per week} = \underset{(0.6)}{-9.47} + \underset{(2.7)}{3.90}\,\text{age} - \underset{(2.0)}{0.07}\,\text{age}^2,\ R^2 = .075 \qquad (2)$$

accidentally spend as much time in school as nonstudents work, this method will yield accurate results. Table 4 gives some indication of the value of this approach as an estimate of schooling time.

Method 3: 40 Times Weekly Pay of Working Youth

This measure used by Schultz (1960) is conceptually partway toward the direct measurement of school hours, since it standardizes for the number of weeks in the school year. If youth attend school at the same rate per week that they would otherwise work and students do not work at all and therefore are not included among working youth, this approach would accurately measure schooling time costs.[10] In table 7, the average number of hours per week worked by male youth are presented in column 1 and should be compared with a standard of 33.3 hours per week (1,300/40).

Method 3a: 40 Times Weekly Pay of Nonstudents

This measure used by Blitz (1962) is a derivative of Schultz's technique. If one assumes that during the school year one attends school at the rate he would otherwise work, then the work efforts of nonstudents are the proper measure of schooling. Looking at the work efforts of all youth (in and out of school), as Schultz was forced to do with his data, should lead to downward-biased estimates of school time. In table 7, average

---

[10] Schultz (1960, p. 574) asserts that ideally his measure should subtract the school year earnings of students. This would be a more accurate measure of earnings differences but a less accurate measure of school time costs, as I show below.

TABLE 8
AGE PROFILES OF VARIOUS ESTIMATES OF SCHOOL HOURS

Age	Direct Measure	Method 1	Method 2	Method 2*a*	Method 3	Method 3*a*
14	1,300	212	370	494	784	1,272
15	1,300	366	576	768	968	1,348
16	1,300	511	764	1,018	1,136	1,420
17	1,300	647	935	1,246	1,284	1,488
18	1,300	773	1,088	1,450	1,412	1,548
19	1,300	890	1,223	1,631	1,520	1,604
20	1,300	998	1,341	1,789	1,608	1,652
21	1,300	1,096	1,443	1,923	1,676	1,696
22	1,300	1,185	1,526	2,035	1,724	1,736

TABLE 9
PERCENTAGE DEVIATION FROM 1,300-HOUR SCHOOL YEAR

Method	High School	College	Study in Which Used
1	−67	−28	Solmon (1971)
2	−49	−2	Becker (1964) and Hanoch (1967)
2*a*	−32	31	Hansen (1963)
3	−20	20	Schultz (1960)
3*a*	6	25	Blitz (1962)

weekly hours of nonstudents are reported in column 2. A quick comparison of the two series suggests major differences only at the lower age groups, where students form a substantial component of all working youth.

Comparisons

In table 8, the estimated schooling hours–age profiles of the various indirect techniques are compared with the measured 1,300 standard. Recall that the direct measures are much firmer for high-school-age youth.

The biases caused by the different techniques can best be summarized, perhaps, by measuring the deviation of each of the estimates from the 1,300-hour annual level for a full high school (ages 14–17) and college (18–21) education. These are presented in table 9. The results indicate that the actual measurement of foregone earnings provides the least useful method of estimating high school time costs, underestimating such costs by almost 70 percent. The Becker-Hanoch assumption at this level also substantially underestimates schooling time, yielding estimates biased downward by 50 percent. The original technique suggested by Schultz and modified by Blitz provides the best estimates at this level. At the college level, the estimates are generally more accurate, with the Becker-Hanoch approach yielding estimates remarkably close to 1,300 hours per year. Three of the studies overestimate schooling costs at this level and substantially so. However, again bear in mind that the 1,300-hour school activity level for college youth is merely an extrapolation of the measured values for high school students

In reviewing the results for high school students, it seems at first ironic that the accuracy of the estimates are roughly inversely related to the time sequence of appearance. This process, however, is not entirely a chance happening, since we are witnessing the increasingly accurate measurement of the wrong variable Differences in work hours between students and nonstudents are not the same as the hours of schooling undertaken by students This paper has presented theoretical reasons why that is so, as well as directly documenting it from measured school hours If anything, at the younger age levels, foregone leisure is much more appropriate as a label for schooling time costs than is foregone earnings

References

Becker, Gary S *Human Capital* New York Columbia Univ Press (for Nat Bur Econ Res), 1964

——— "The Allocation of Time and Goods over Time " Unpublished manuscript, 1967 (*a*)

——— *Human Capital and the Personal Distribution of Income An Analytical Approach.* W S Woytinsky Lecture no 1 Ann Arbor Inst. Public Admin and Dept Econ , Univ Michigan, 1967 (*b*)

Ben-Porath, Yoram "The Production of Human Capital and the Life Cycle of Earnings " *J P E* 75 (August 1967) 352–65

Blitz, Rudolph, C "The Nation's Educational Outlay " In *Economics of Higher Education,* edited by Selma J Mushkin Washington Government Printing Office, 1962

Coleman, James S , et al *Equality of Educational Opportunity* Vols 1 and 2 Washington Government Printing Office, 1966.

Educational Commission of the States. "State Action Extending the School Year " *Compact* 4 (December 1970) 38–44

Gronau, Reuben. "The Intrafamily Allocation of Time The Value of the Housewives' Time." Mimeographed, November 1971.

Gustman, A L "Human Capital, Capital Theory, and Occupational Earnings." Unpublished Ph D dissertation, Univ. Michigan, 1969.

Hanoch, Giora. "An Economic Analysis of Earnings and Schooling." *J. Human Resources* 2 (Summer 1967). 310–29.

Hansen, W Lee "Total and Private Rates of Return to Investment in Schooling." *J P E* 71 (April 1963)· 128–40

Lindsay C. M "Measuring Human Capital Returns." *J.P.E.* 79 (December 1971)· 1195–1215

Marwardt, Frederick, and Sikkink, Donald E. "Student Preparation Time." *Improving College and Univ. Teachers* 18 (Autumn 1970). 308–9.

National Education Association, Research Division. *NEA Research Bulletin* (March 1970), p. 27

Parsons, Donald O. "Intergenerational Wealth Transfers and the Educational Decisions of Male Youth." Mimeographed, 1972.

Schultz, Theodore W "Capital Formation by Education." *J.P.E.* 68 (December 1960) 571–83

——— "Rise in the Capital Stock Represented by Education in the United States, 1900–57 " In *Economics of Higher Education*, edited by Selma J. Mushkin. Washington Government Printing Office, 1962

Solmon, Lewis C "Capital Formation by Expenditures on Education in 1960." *J P.E* 79 (December 1971) 1412–17.

Stafford, Frank. "The Income and Consumption of Graduate Students." Unpublished Ph D dissertation, Univ Chicago, 1968.

Advertising and Profitability: A Reappraisal

Harry Bloch

University of Manitoba

Previous empirical studies of the relationship between profitability, market structure, and advertising intensity have consistently found a strong positive relationship between profit rates and advertising intensity. The profit rates used in these studies are based on treating advertising expenditures as a current expense. Much evidence exists that advertising has a long-lasting effect on sales, so that the treatment of advertising as a current expense is inappropriate. In this paper, corrected profit rates are calculated by treating advertising expenditures as an investment in a capital asset. The corrected profit rates are then used in a multivariate regression analysis of the relationship between profitability, market structure, and advertising intensity. No significant relationship between corrected profit rates and advertising intensity is found in the regressions. This finding suggests that the positive relationships between uncorrected profit rates and advertising intensity in previous studies are due solely to the inappropriate expensing of advertising.

Beginning with a study by Comanor and Wilson (1967), there have been a number of studies using multivariate regression analysis to investigate the relationship between profit performance, advertising intensity, and market structure. The general method of the studies is to use profit rates equal to reported profit divided by reported net worth as the dependent variable and measures of advertising intensity and market structure as independent variables. The conclusion which is reached consistently is that there is a strong positive relationship between profit rates and advertising intensity.

The measures of profit rate used in the studies are subject to error due to inappropriate accounting treatment of advertising in the calculation of reported profit and net worth. Advertising expenditures are treated

I am indebted to Lester Telser, Marc Nerlove, Dale Orr, Maurice Levi, and Leonard Weiss for constructive criticism and advice.

as a current expense, even though they are generally believed to have a long-lasting impact on sales and should therefore be treated as an investment The presence of errors in the measures of profit rate used can cause the results of regressions of the above type to be misleading

Weiss (1969), finds that adjusting the profit rates of a sample of industries to correct for the expensing of advertising does not much change the strong positive relationship between profit rates and advertising intensity (pp 428–29) However, there is evidence that his adjustments are inaccurate We argue below that the questionable accuracy of the adjustments is due to aggregation problems encountered in the use of industry data We find that when corrections for the expensing of advertising are made to the profit rates of a sample of firms, the strong positive relationship between profit rates and advertising intensity disappears This finding supports the view that the strong positive relationships found in previously reported studies are due merely to the expensing of advertising

I. The Effects of Expensing Advertising

Studies which demonstrate that advertising has a long-lasting effect on sales have been done by Nerlove and Waugh (1961), Telser (1962), Palda (1964), Lambin (1970), and Peles (1971) Yet, normal accounting practice is to treat advertising expenditures as a current expense in the calculation of reported profit and net worth The assets on the balance sheets of firms exclude the value of undepreciated advertising expenditures, so that the reported net worth of the firms is understated Also, when firms have current expenditures on advertising which differ from the amount of depreciation of their advertising asset, the reported profit of the firms is misstated

The effect that the expensing of advertising has on a measure of profit rate equal to reported profit divided by reported net worth can be stated in terms of the difference between measured and "true" profit rates The "true" profit rate is defined for our purposes as the rate which would prevail if advertising expenditures were treated as an investment and capitalized over their full economic lives[1] Understatement of reported net worth causes the measured profit rate to be above the true profit rate However, this effect can be partially or totally offset by positive net investment in advertising, which causes the measured rate to be below the true rate[2]

---

[1] The "true" profit rate which we use is not corrected for all inappropriate accounting In addition to advertising, expenditures on research and development, product promotion, and on-the-job training are treated as current expenses, even though they can be expected to have long-lasting effects No explicit treatment is given to the inappropriate accounting of these additional classes of expenditure because of insufficient data to calculate the effect on profit rates Thus, our true profit rate is still subject to error, but our purpose is merely to remove spurious correlation between advertising and profitability due to the inappropriate expensing of advertising expenditures.

The general method of studies which have investigated the relationship between profit performance, advertising intensity, and market structure is to calculate regressions of the following type

$$\pi/E = a_0 + \sum_{i=1}^{n} a_i X_i + e, \tag{1}$$

where π/E is the measured profit rate equal to reported profit divided by reported net worth, the a's are estimated coefficients, the X_i's are measures of advertising intensity and market structure, and e is a residual term The above discussion suggests that the measured profit rate used is composed of a true profit rate component and an error term component The presence of an error term in the dependent variable can cause regression coefficients to reflect inaccurately the relationships between the independent variables and profitability[3] In particular, the coefficient of advertising intensity is likely to be affected, because advertising intensity can be expected to be correlated with the amount of advertising asset and net investment in advertising Telser (1969) has suggested that advertising intensity may serve as a proxy for omitted intangible capital in the form of advertising, so that the coefficient of advertising intensity in regressions of type (1) is biased upward (pp 122–23) A positive coefficient of advertising intensity in regressions of type (1) can indicate a positive relationship between advertising intensity and the error term, rather than indicating that advertising increases profitability To determine the relationship between advertising and true profitability, it is necessary to correct for the expensing of advertising in the calculation of measured profit rates

II. Weiss's Study

Weiss (1969) investigates the effect that the expensing of advertising has on coefficients in regressions of type (1) He does this by comparing coefficients obtained from regressions of type (1) to coefficients obtained from regressions of the following type

$$\pi^*/E^* = a_0' + \sum_{i=1}^{n} a_i' X_i + e', \tag{2}$$

---

[2] If we assume that advertising expenditures grow at a constant rate over time, we can solve for the difference between measured and true profit rates in terms of the rate of growth of advertising expenditures, the rate of depreciation of advertising expenditures, and the level of the measured profit rate. A derivation of this relationship given by Telser (1968, p 167) and by Weiss (1969, pp 421–23)

[3] For a discussion of the effects of errors in the measurement of the dependent variable on the coefficients in classical least-squares regressions, see Johnston (1963, pp. 148–62) or Goldberger (1964, pp. 282–84).

where π^*/E^* is an estimate of the true profit rate derived from the measured profit rate, π/E, by treating advertising expenditures as an investment rather than a current expense, and the X_i's are the same independent variables used in the corresponding regression of type (1). He finds the regression coefficients in (2) are close to the corresponding coefficients in (1) The coefficients of advertising intensity are only moderately lower in the regressions of type (2), and they are still greater than zero at the 1 percent significance level using a one-tailed t-test These findings suggest that the expensing of advertising does not affect the relationship between profit rates and advertising intensity, and that the coefficient of advertising intensity in regressions of type (1) provides an accurate indication of the relationship between advertising and profitability

In calculating estimates of true profit rates for use in regressions of type (2) it is necessary to make assumptions regarding the manner in which advertising expenditures depreciate over time In studies which have attempted to estimate the pattern of depreciation of advertising expenditures, the estimated pattern varies substantially with the type of product advertised and the method used in estimation[4] This disparate evidence on depreciation patterns and problems of aggregation which are explained below make it important to check the accuracy of estimated true profit rates

A test of the accuracy of the estimate of true profit rates used in regressions of type (2) is possible by making use of the relationship between measured and true profit rates Weiss's estimates of true profit rates are derived from calculations of the following form $\pi^*|E^* = (\pi + \Delta Z)/(E + Z)$, where π^*/E^* = the true profit rate for a period of time, π = *the reported profit for the period*, ΔZ = *net investment in* advertising during the period, E = reported net worth at the end of the period, Z = accumulated undepreciated advertising expenditures at the end of the period[5] Using this expression, the measured profit rate, π/E, can be expressed as $\pi/E = (\pi^*/E^*)(1 + Z/E) - \Delta Z/E$ If the relationship between the X_i's and true profit rates is properly specified in regres-

---

[4] By calculating regressions involving current and lagged market shares, Telser (1962) estimated that the advertising expenditures of cigarette firms depreciated at a constant annual rate varying between 15 and 20 percent (p 498) By calculating regressions involving current and lagged absolute sales, Palda (1964) estimated that the advertising expenditures for Lydia Pinkham Vegetable Compound depreciated at a constant annual rate of 33 percent (pp 60–68) By calculating regressions of similar form to Telser's Lambin (1970) estimated that the advertising expenditures for an unspecified consumer durable depreciated at a constant annual rate between 29 and 60 percent (p 475).

[5] Weiss's formula for calculating true profit rates is as follows:

$$\frac{\pi_t^*}{E_t^*} = \frac{\pi_t + A_t - (1 - \lambda) \sum_{i=0}^{4} (\lambda^i A_{t-i}) - \lambda^5 A_{t-5}}{E_t + \sum_{i=0}^{4} \lambda^{i+1} A_{t-i}},$$

where A_t is advertising expenditures in year t and λ is one minus the annual rate of depreciation of advertising expenditures The values of Z and ΔZ implied by this formula sions of type (2), the relationship between measured profit rates and the same X_i's is properly specified in regressions of the following type

$$\pi/E = a_0''(1 + Z/E) + \sum_{i=1}^{n} a_i''(X_i(1 + Z/E)) + a_{n+1}(\Delta Z/E) + \epsilon'' \qquad (3)$$

The dependent variable in regressions of type (3) is the same as in regressions of type (1), but the independent variables are altered to correct for the expensing of advertising. If altering the independent variables in a regression of type (3) reduces errors associated with the expensing of advertising, the corrected R^2 for the regression should be higher than the corrected R^2 for a corresponding regression of type (1). Thus, a comparison of the corrected R^2 provides a test of the accuracy of the adjustments made in calculating estimates of the true profit rates.[6] A further test of the accuracy of the adjustments is given by the value of the coefficient a_{n+1} in the regressions of type (3). The assumed value of this coefficient in Weiss's calculation of estimates of the true profit rates is -1, so that a_{n+1} should have a value of -1.

We calculate regressions of type (3) with the data on measured profit rates, advertising intensity, and market structure which Weiss uses in his regressions and with values of Z/E and $\Delta Z/E$ implied by his calculations of true profit rates. The results of these regressions are listed on lines 5 and 6 of table 1. Also listed in table 1 are the results presented by Weiss for the corresponding regressions of type (2), listed on lines 3 and 4, and the corresponding regressions of type (1), listed on lines 1 and 2. The corrected R^2 for each of the regressions of type (3) is below the corrected R^2 for the corresponding regression of type (1), and the values of the coefficient of $\Delta Z/E$ differ substantially from -1.[7] Thus, Weiss's estimates of true profit rates do not pass the test of accuracy established above.

---

are as follows

$$\Delta Z = A_t - (1 - \lambda) \sum_{i=0}^{4} (\lambda^i A_{t-i}) - \lambda^5 A_{t-5},$$
$$Z = \sum_{i=0}^{4} \lambda^{i+1} A_{t-i}$$

[6] If measured profit rates are subject to error due to inappropriate expensing of advertising, regressions of type (1) are improperly specified. The independent variables are altered in regressions of type (3) to correct for this improper specification. The improper specification will not be corrected if inaccurate values of Z/E and $\Delta Z/E$ are used in regressions of type (3). The size of the residual variances provides a test for an improvement in specification in going from (1) to (3). If a regression of type (3) is properly specified while the corresponding regression of type (1) is improperly specified, the residual variance for the regression of type (3) should be smaller according to Theil (1971, p. 543). As the total variance is the same in corresponding regressions of types (1) and (3), the corrected R^2 calculated according to the formula in n. 7 below is inversely related to the residual variance.

[7] The corrected R^2 listed on lines 1 through 4 of table 1 are not those given by Weiss (1969, p. 429). Rather, to obtain equivalent values of R^2 for all the regressions in table 1, all the R^2 were calculated using the following conventional formula for corrected R^2: $\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1)$, where n is the number of observations and k is the number of explanatory variables.

TABLE 1

REGRESSIONS RELATING π/E AND π^*/E^* TO VARIOUS EXPLANATORY VARIABLES FOR 37 CONSUMER-GOODS INDUSTRIES (t-RATIOS IN PARENTHESES)

Regression	Intercept	Advertising-to-Sales Ratio	High Advertising Dummy	Capital Requirements (logs)	Demand Growth (logs)	Regional Industry Dummy	Ratio of Advertising Investment to Net Worth ($\Delta Z/E$)	R^2	Corrected R^2
Type (1) from Weiss (1969) 1963–64 π/E dependent									
1. Eq (3) of table 6	−0 0283	0 4496		0 0091	0 0295	−0 0071		35	27
		(2 75)		(2 36)	(1 19)	(0 40)			
2. Eq (4) of table 6	0 0227		0 0411	0 0052	0 0160	−0 0086		39	31
			(3 20)	(1 33)	(1 11)	(0 50)			
Type (2) from Weiss (1969) 1963–64 π^*/E^* dependent									
3. Eq (5) of table 6	−0 0155	0 3859		0 0077	0 0371	−0 0096		32	.24
		(2 40)		(2 03)	(1 52)	(0.55)			
4. Eq (6) of table 6	0 0380		0 0382	0 0041	0 0332	−0 0112		38	.30
			(3 07)	(1 07)	(1 42)	(0 67)			
Type (3): 1963–64 π/E dependent									
5 Corresponding to eqq (1) and (3)	−0 0236	0 0784		0 0085	0 0299	−0 0061	0 2385	32	.21
		(0 41)		(2 30)	(1 27)	(0 39)	(0 42)		
6 Corresponding to eqq (2) and (4)	0 0032		0 0219	0 0062	0 0260	−0 0083	−0 0300	.38	.27
			(1 68)	(1 63)	(1 16)	(0 56)	(0 07)		

The data used in calculating the regressions in table 1 consist of observations on 37 U S. consumer-goods industries. Weiss measures the industry level value of Z by the sum of current and past advertising expenditures of all firms in an industry minus the accumulated depreciation of these expenditures. He measures the industry-level value of ΔZ by the sum of current advertising expenditures by all firms in an industry minus current depreciation of Z. These procedures are equivalent to taking the sum, for all firms in an industry, of each firm's Z and ΔZ, valued as the firm's own advertising expenditures minus depreciation. Values of Z and ΔZ calculated in this manner tend to overstate an industry's advertising asset and net investment in advertising whenever firm advertising expenditures have long-lasting negative effects on the demand facing other firms in the industry [8] The limited amount of evidence available indicates that these negative effects are substantial, which implies that there is a serious aggregation problem in calculating industry-level advertising assets and net investments in advertising [9] Improper aggregation of firm advertising expenditures could explain the apparent inaccuracy of Weiss's estimates of industry true profit rates [10] In any event, the possibilities of improper aggregation add to the difficulties of obtaining accurate estimates of true profit rates at the industry level

III. Results Using Firm Data

The aggregation problems encountered in obtaining accurate estimates of true profit rates for industries can be lessened by using estimates of true profit rates for firms [11] Firm profit-rate data are used in three studies of the relationship between profit rates, advertising intensity, and market structure Federal Trade Commission (1969), Imel and Helmberger (1971), and Shepherd (1972). However, a sufficient amount of data on advertising expenditures over time, to allow reasonable calculations of values of Z and ΔZ, are available only for the firms in the study by the Federal Trade Commission, and even so for only some of the firms Our investigation of the effect of inappropriate expensing of advertising on the coefficients in regressions of type (1) is therefore based on a subsample of the FTC firms

---

[8] When a firm's advertising has long-lasting negative effects on its competitors, the firm's advertising creates a negative asset for the competitors as well as a positive asset for the firm itself Thus means that the addition to an industry's advertising assets from each firm's advertising is smaller than the addition to the firm's own advertising asset Thus, when the values of Z and ΔZ for the firms in an industry are appropriately measured, and when each firm's advertising has negative effects on its competitors, values of Z and ΔZ for the industry calculated as the simple sum of the values of Z and ΔZ for firms in the industry overstate the industry's advertising asset and net investment in advertising

[9] Telser (1962) supplies evidence on the negative effects of cigarette-firm advertising on other cigarette firms (pp 483–87) Peles (1971) supplies similar evidence for the advertising of cigarette, beer, and automobile firms (pp 39–43, 72–73) Finally, Lambin (1970) supplies evidence of the negative effects of competitors' advertising on the demand for an unspecified consumer durable (pp. 479–80)

[10] Another explanation of the apparent inaccuracy is that it is due to assuming the wrong rate of depreciation of advertising expenditures Attempts to improve the accuracy of the estimates of true profit rates by using alternative assumptions about the rate of depreciation met with no success.

[11] Even when a firm's advertising has long-lasting negative effects on its competitors, the value of the advertising asset created for the firm will usually include only the positive effects on the firm's own future profits. An exception is where the firm's advertising leads its competitors to increase their advertising, and this induced advertising has a long-lasting negative effect on the first firm's future profits.

In the FTC study, regressions of type (1) are calculated using data for 97 U S food manufacturing firms The profit rates used in the regressions are simple averages of annual profit rates for each of the 5 years from 1949 to 1953 Two profit rates are used a profit rate P_1 which is equal to reported net income after taxes plus interest on long-term debt divided by reported net worth plus long-term debt, and a profit rate P_2 which is equal to reported net income after taxes divided by reported net worth For each profit rate, linear and nonlinear regressions of type (1) in both unweighted and weighted form are calculated, using as independent variables advertising-to-sales ratios, concentration ratios, relative market shares, percentage changes in industry demand, one over the log of firm assets, and alternatively, three-digit-, four-digit-, and five-digit-firm diversification ratios For the purpose of investigating the effect of the expensing of advertising on the regression coefficients, we use unweighted linear regressions with five-digit diversification ratios[12] The results listed by the FTC for the regressions in this form are repeated here on lines 1 and 2 of table 2

The data used in the calculation of the FTC regressions are listed in appendix C to their study To check if the data are consistent with the regression results listed on lines 1 and 2 of table 2, regressions of identical form are calculated using the data The results of these latter regressions are listed on lines 3 and 4 of the table. While the results of corresponding regressions of identical form differ, the strong positive relationship between profit rates and advertising-to-sales ratios is not much affected

The advertising-to-sales ratios used by the FTC are found by taking averages of the advertising-to-sales ratios for all the industries in which a firm operates Because the implied measures of advertising expenditure do not reflect the actual expenditures of the firms, they are not appropriate

---

[12] The choice of linear rather than nonlinear regressions was based on a desire to avoid the multicolinearity which arises when several different powers of the same explanatory variable are used in a regression. Regressions with five-digit-firm diversification ratios were chosen because these regressions had R^2s which were uniformly higher than the R^2s for the corresponding regression with a three- or four-digit-firm diversification ratio. Finally, unweighted regressions were chosen because heteroscedasticity was not present in the residuals of the regressions using the 40-firm subsample on which the conclusions of this study are based

TABLE 2

Regressions Relating Measured Profit Rates to Various Explanatory Variables for 97 Food Manufacturing Firms and Subsamples of 57 and 40 Firms (*t*-Ratios in Parentheses)

Regression	Profit Rate	Intercept	Four-Firm Concentration Ratio (%)	Relative Market Share (%)	Advertising-to-Sales Ratio (%)	Change in Industry Demand (%)	Five-Digit-Firm-Diversification Ratio (%)	$\frac{1}{\text{Log Assets}}$	R^2	Corrected R^2
From FTC (1969) for full 97-firm sample										
1. Eq. (3a) of table 3-2	P_1	8.56	0.039	0.074	1.07	0.014	−0.069	−5.34	.410	.371
			(1.06)	(2.05)	(2.47)	(1.01)	(3.04)	(1.41)		
2. Eq. (3b) of table 3-2	P_2	13.60	0.049	0.053	1.16	0.020	−0.093	−12.3	.397	.357
			(1.10)	(1.22)	(2.20)	(1.17)	(3.37)	(2.67)		
For full 97-firm sample:										
3.	P_1	7.18	0.040	0.084	1.16	−0.005	−0.056	−3.69	.420	.381
			(1.08)	(2.82)	(2.73)	(0.32)	(2.49)	(1.03)		
4.	P_2	12.18	0.050	0.065	1.25	0.001	−0.081	−10.5	.396	.356
			(1.11)	(1.79)	(2.41)	(0.05)	(2.91)	(2.39)		
For 57-firm subsample:										
5.	P_1	4.10	0.083	0.107	0.503	−0.019	−0.054	−1.09	.329	.249
			(1.21)	(1.82)	(0.65)	(0.86)	(1.51)	(0.18)		
6.	P_2	10.35	0.087	0.087	0.844	−0.014	−0.080	−9.71	.325	.244
			(0.99)	(1.16)	(0.86)	(0.52)	(1.78)	(1.27)		
For 40-firm subsample:										
7.	P_1	3.98	0.027	0.086	1.67	0.010	−0.042	0.280	.629	.562
			(0.69)	(2.56)	(3.33)	(0.50)	(1.35)	(0.04)		
8.	P_2	6.68	0.039	0.070	1.44	0.014	−0.056	−1.85		
			(0.93)	(1.93)	(2.67)	(0.67)	(1.67)	(0.25)	.585	.510

for calculating estimates of true profit rates for the firms. For 40 out of the 97 firms in the FTC regression sample, data on firm advertising expenditures in six major media were available for the period 1936 through 1950 inclusive, and for the individual years, 1950, 1951, 1952, and 1953.[13] Our investigation for the effects of the expensing of advertising on the coefficient of advertising intensity in regressions of type (1) is therefore based on the 40-firm subsample. To check whether use of the 40-firm subsample distorts the results of the investigation, regressions in the same form as the regressions on lines 1–4 of table 2 are calculated for both the 40-firm subsample of included firms and the 57-firm subsample of excluded firms. The results of the 57-firm regressions are listed on lines 5 and 6 of table 2, and the results of the 40-firm regressions are listed on lines 7 and 8. The coefficients of the advertising-to-sales ratio are higher in the 40-firm regressions than in either the 57-firm regressions or the 97-firm regressions.

The data on advertising expenditures in six major media for the 40-firm subsample are used to calculate measures of Z and ΔZ by firm for each of the years 1950 through 1953.[14] These calculations assume a 5 percent per annum rate of depreciation of advertising capital. This rate of depreciation is chosen because it implies values of Z capable of explaining variations in the relationship between the market value and book value of the stock of firms in the regression sample.[15] The calculated measures

---

[13] The data on advertising expenditures in the six media for the years 1951, 1952, and 1953 are given in the October 29, 1954 issue of *Printer's Ink*, while the data on advertising expenditures for the year 1950 and for the period 1936–50 inclusive are given in the November 23, 1951 issue. The six media for which advertising expenditures are included are network television, network radio, magazines, newspapers, national farm papers, and Sunday magazine supplements.

[14] The first step in calculating measures of Z and ΔZ was to solve for values of r, the compounded ratio of advertising expenditures in year $t - 1$ to advertising expenditures in year t. These values were found using the following formula

$$\sum_{t=1936}^{1950} A_t = \frac{[(1 - r)^{15}]A_{1950}}{1 - r},$$

where A_t is advertising expenditures in year t. Using the derived values of r, measures of Z and ΔZ were then calculated from the following formulas, where λ is one minus the annual rate of depreciation of advertising expenditures. $Z_{1950} = [\lambda/(1 - \lambda r)]A_{1950}$, $Z_t = \lambda(A_t + Z_{t-1})$ for $t = 1951$, 1952, and 1953, $\Delta Z_t = A_t - [(1 - \lambda)/\lambda)]Z_t$ for $t = 1950$, 1951, 1952, and 1953.

[15] The treatment of advertising as a current expense means that the amount of undepreciated advertising expenditures is not included as an asset on balance sheets, so that the value of a firm's advertising asset is not reflected in the book value of the firm's stock. The value of the firm's advertising asset does, however, tend to be reflected in the market value of the firm's stock. Thus, accurate measures of Z for firms should be highly correlated with the difference between the book value and market value of the firms' stock. We calculate regressions of the following form for 37 of the firms in the 40-firm subsample: $\log (M/E) = b_1 + b_2 \log [(E + Z)/E] + b_3 \log [\pi/(E + Z)] + e$, where M is the market value of the firm's stock. The values of Z used in the regressions are calculated using, alternatively, 5, 10, 20, and 33 percent annual rates of depreciation of advertising expenditures. The R^2s for the regressions decreased continually from .679 for

of Z and ΔZ are used to find estimates of the two average true profit rates, P_1^* and P_2^*, corresponding to the measured profit rates, P_1 and P_2. Because advertising expenditure data are not available for the year 1949, the period over which the average profit rates are calculated is limited to the 4 years 1950 through 1953, rather than the 5 years 1949 through 1953 used in the FTC study. The formulas used in finding the estimates of true profit rates are equivalent to those used by Weiss, except for the treatment of net investment in advertising.[16] Weiss adds the full amount of net investment in advertising to the numerator of his estimates of true profit rates. We add only the portion of net investment in advertising which is not offset by a reduction in taxes. Our purpose is to have estimates of true profit rates which reflect what profits would have been if advertising had been treated as an investment rather than expensed.[17] Our estimates of P_1^* and P_2^* for the 40 firms are listed in table A1, along with the corresponding values of P_1 and P_2.

---

the regression with the 5 percent depreciation rate to .674 for the regression with the 33 percent depreciation rate. The t-ratio for the coefficient b_2 decreased continually from 4.79 for the 5 percent regression to 1.62 for the 33 percent regression.

[16] The average profit rates were calculated using the following formulas:

$$P_1 = \left(\sum_{t=1950}^{1953} \pi_{1t}\right) \Big/ \left(\sum_{t=1950}^{1953} E_{1t}\right),$$

$$P_2 = \left(\sum_{t=1950}^{1953} \pi_{2t}\right) \Big/ \left(\sum_{t=1950}^{1953} E_{2t}\right),$$

$$P_1^* = \frac{\sum_{t=1950}^{1953} [\pi_{1t} + (1 - m_t)\,\Delta Z_t]}{\sum_{t=1950}^{1953} (E_{1t} + Z_t)},$$

$$P_2^* = \frac{\sum_{t=1950}^{1953} [\pi_{2t} + (1 - m_t)\,\Delta Z_t]}{\sum_{t=1950}^{1953} (E_{2t} + Z_t)},$$

where π_{1t} = reported net income after taxes plus interest on long-term debt in year t, E_{1t} = reported net worth plus long-term debt in year t, π_{2t} = reported net income after taxes in year t, E_{2t} = reported net worth in year t, and m_t = the marginal corporate income tax rate in year t. The values of m_t used in the calculations include all income tax surcharges and excess profits taxes in effect during the years 1950 through 1953. To allow for the fact that the fiscal years of the firms ended at various times between June 30 of year t and June 29 of year $t + 1$, the values of m_t were calculated by taking a simple average of the marginal corporate income tax rates in effect at the end of each month from June of year t to May of year $t + 1$, as indicated by the Internal Revenue's Service's *Statistics of Income* (Washington: Government Printing Office), for the years 1950 through 1954: $m_{1950} = 0.57$, $m_{1951} = 0.80$, $m_{1952} = 0.82$, and $m_{1953} = 0.79$. Values of reported net income after taxes, interest on long-term debt, and reported net worth were taken from *Moody's Industrials* for the years 1950 through 1954.

[17] In the numerator of the estimates of true profit rates ΔZ_t is multiplied by $(1 - m_t)$ because the profit rates are calculated on an after-tax basis. The expensing of advertising results in reported profits before taxes being lower by one dollar for each dollar of net investment in advertising. When reported profits are lower, so are income taxes. Hence, the expensing of advertising has the effect of lowering reported profits after taxes by the amount of net investment in advertising times one minus the applicable income tax rate.

The 4-year average measured profit rates, P_1 and P_2, are used in calculating regressions of type (1) identical with those on lines 7 and 8 of table 2, except for the slightly shorter period over which the profit rates are averaged. The results of these regressions are listed on lines 1 and 2 of table 3. On lines 3 and 4 of table 3, are listed the results of regressions of type (2) which have the estimated true profit rate, P_1^* or P_2^*, substituted for the corresponding measured profit rate as the dependent variable. The coefficient of the advertising-to-sales ratio is substantially lower in the regressions of type (2) than in the corresponding regression of type (1). This result provides evidence that the coefficient of the advertising-to-sales ratio in regressions of type (1) overstates the relationship between advertising and profitability. Furthermore, the fact that the coefficients of the advertising-to-sales ratio in the regressions of type (2) are not greater than zero at the 5 percent significance level using a one-tailed t-test indicates that there is actually no measurable relationship between advertising and profitability.

The reliability of the above evidence depends on the accuracy of the estimates of true profit rates used in the regressions of type (2). According to the analysis of section II above, the accuracy of the estimates can be checked by calculating regressions of type (3) corresponding to the regressions of types (1) and (2). The results of such regressions of type (3) are listed in lines 5 and 6 of table 3. The corrected R^2 for these regressions of type (3) are higher than the corrected R^2 for the corresponding regressions of type (1). Also, the coefficients of $\Delta Z/E$ have values close to those assumed in the calculation of the estimates of true profit rates, although these coefficients do not differ from zero at the 5 percent significance level using a two-tailed t-test.[18] These results conform reasonably well to what would be expected from accurate estimates of true profit rates.

The advertising-to-sales ratios used in the regressions of table 3 are computed by the FTC from industry average advertising-to-sales ratios. These computed advertising-to-sales ratios do not reflect the actual ratios of advertising expenditure to sales for the individual firms in the regression sample, and may therefore be viewed as inappropriate. An alternative measure of the advertising-to-sales ratio is computed by dividing firm advertising expenditures in six major media for the years 1950 through 1953 inclusive by firm sales for the same period.[19] The alternative advertising-to-sales ratios are used in calculating regressions of types (1)–(3) identical with those in table 3, except for the change in

[18] The average value of $(1 - m_t)$ used in calculating true profit rates according to n. 16 above is 0.255. In regressions of type (3), this implies a coefficient $\Delta Z/E$ of -0.255.

[19] The data on advertising expenditures in the six media was taken from *Printer's Ink* issues of October 29, 1954 and November 23, 1951, and the data on net sales after excise taxes was taken from *Moody's Industrials* for the years 1950 through 1954.

the measure of the advertising-to-sales ratio The results of these regressions, which are listed in table 4, do not differ in any consistent way from the results of the regressions listed in table 3, except for the somewhat lower coefficients of the advertising-to-sales ratio in table 4.

The regressions in tables 3 and 4 are calculated based on the assumption that advertising expenditures depreciate at a rate of 5 percent per annum Given that there is uncertainty regarding the actual depreciation rate, it is important to check the sensitivity of our results to changes in the assumed depreciation rate In his study, Weiss assumes advertising expenditures depreciate at 33 3 percent per annum We calculate values of Z, ΔZ, P_1^*, and P_2^* for the subsample of 40 food manufacturing firms assuming a 33 3 percent advertising depreciation rate and then use these values to calculate regressions similar to those in table 4 The results of these regressions are listed in table 5 The regressions are numbered with primes, so that each regression corresponds to the unprimed regression of the same number in table 4

The higher assumed advertising depreciation rate in the regressions of table 5 is associated with substantially higher coefficients for the advertising-to-sales ratio However, these coefficients are not greater than zero at the 5 percent significance level using a one-tailed *t*-test Also the corrected R^2 for the regressions of type (3) in table 5 are below the corrected R^2 for the corresponding regressions in table 4 In section II above, we show that the corrected R^2 in regressions of type (3) provides a test of the accuracy of the adjustments made in calculating true profit rates Hence, we have evidence that calculations of true profit rates based on an assumed 33.3 percent annual advertising depreciation rate are less accurate than calculations based on an assumed 5 percent rate, at least for the firms in the 40-firm subsample

In both table 3 and table 4 there are substantial differences between the coefficient of the advertising-to-sales ratio in the regressions of type (1) and the coefficient of the advertising-to-sales ratio in the corresponding regressions of type (2) or (3) This finding provides evidence that the coefficient of the advertising-to-sales ratio in regressions of type (1) overstates the relationship between advertising and profitability While a strong positive relationship between profit rates and advertising intensity is found in the regressions of type (1), the relationship disappears in the regressions of types (2) and (3) when corrections are made for the expensing of advertising. Switching from the use of measured profit rates as the dependent variable in regressions of type (1) to the use of true profit rates as the dependent variable in regressions of type (2) affects the coefficient of no other variable as strongly or persistently as it affects the coefficient of the advertising-to-sales ratio. Therefore, it is appropriate to view the advertising-to-sales ratio in the regressions of type (1) as a proxy for errors in measured profit rates resulting from the expensing of advertising.

TABLE 3

Regressions Relating Measured and True Profit Rates to Various Explanatory Variables for 40 Food Manufacturing Firms, Using Advertising-to-Sales Ratios Computed by the FTC (t-Ratios in Parentheses)

Regression	Profit Rate	Intercept	Four-Firm Concentration Ratio (%)	Relative Market Share (%)	Advertising-to-Sales Ratio (%)	Change in Industry Demand (%)	Five-Digit Firm Diversification Ratio (%)	Ratio of Advertising Investment to Net Worth ($\Delta Z/E$) (%)	$\frac{1}{\text{Log Assets}}$	R^2	Corrected R^2
Type (1) 1950–53 average measured profit rate dependent											
1.	P_1	2.45	0.020	0.078	1.45	0.003	−0.028		3.87	.628	.560
			(0.59)	(2.74)	(3.39)	(0.17)	(1.06)		(0.66)		
2.	P_2	5.29	0.026	0.065	1.23	0.007	−0.041		1.95	.585	.510
			(0.75)	(2.15)	(2.75)	(0.39)	(1.44)		(0.32)		
Type (2)· 1950–53 average "true" profit rate dependent.											
3.	P_1^*	3.33	0.030	0.064	0.202	0.001	−0.025		2.96	.576	.500
			(1.33)	(3.34)	(0.71)	(0.06)	(1.39)		(0.76)		
4.	P_2^*	5.03	0.034	0.058	0.064	0.002	−0.031		1.46	.536	.452
			(1.43)	(2.87)	(0.21)	(0.14)	(1.62)		(0.35)		
Type (3). 1950–53 average measured profit rate dependent											
5	P_1	3.43	0.023	0.066	0.277	0.0003	−0.025	−0.172	2.52	.703	.638
			(1.05)	(3.24)	(1.14)	(0.03)	(1.38)	(0.76)	(0.62)		
6	P_2	5.09	0.028	0.062	0.094	0.002	−0.031	−0.102	0.759	.631	.550
			(1.19)	(2.73)	(0.37)	(0.19)	(1.65)	(0.49)	(0.17)		

TABLE 4

REGRESSIONS RELATING MEASURED AND TRUE PROFIT RATES TO VARIOUS EXPLANATORY VARIABLES FOR 40 FOOD MANUFACTURING FIRMS, USING ADVERTISING-TO-SALES RATIO EQUAL TO 1950–53 ADVERTISING EXPENDITURES DIVIDED BY 1950–53 SALES (t-RATIOS IN PARENTHESES)

Regression	Profit Rate	Intercept	Four-Firm Concentration Ratio (%)	Relative Market Share (%)	Advertising-to-Sales Ratio (%)	Change in Industry Demand (%)	Five-Digit-Firm-Diversification Ratio (%)	Ratio of Advertising Investment to Net Worth ($\Delta Z/E$) (%)	$\frac{1}{\text{Log Assets}}$	R^2	Corrected R^2
Type (1). 1950–53 average measured profit rate dependent											
1. .	P_1	4.17	0.050	0.073	0.881	0.022	−0.043		0.211	569	.491
			(1.50)	(2.39)	(2.32)	(1.27)	(1.51)		(0.03)		
2.	P_2	7.01	0.047	0.062	0.897	0.024	−0.053		−1.91	564	.486
			(1.43)	(2.01)	(2.37)	(1.37)	(1.86)		(0.29)		
Type (2): 1950–53 average true profit rate dependent											
3. .	P^*_1	3.41	0.037	0.062	0.030	0.003	−0.027		2.93	570	492
			(1.78)	(3.23)	(0.13)	(0.29)	(1.53)		(0.71)		
4.	P^*_2	5.02	0.036	0.057	−0.014	0.002	−0.032		1.57	536	451
			(1.68)	(2.84)	(0.06)	(0.21)	(1.69)		(0.36)		
Type (3). 1950–53 average measured profit rate dependent											
5.	P_1	3.61	0.034	0.062	0.018	0.005	−0.030	−0.108	2.20	691	.624
			(1.57)	(3.02)	(0.06)	(0.49)	(1.65)	(0.34)	(0.53)		
6.	P_2	4.98	0.037	0.060	−0.154	0.004	−0.034	0.010	0.760	632	.552
			(1.61)	(2.68)	(0.50)	(0.41)	(1.83)	(0.04)	(0.18)		

TABLE 5

REGRESSIONS RELATING MEASURED AND TRUE PROFIT RATES TO VARIOUS EXPLANATORY VARIABLES FOR 40 FOOD MANUFACTURING FIRMS, USING ADVERTISING-TO-SALES RATIO EQUAL TO 1950–53 ADVERTISING EXPENDITURES DIVIDED BY 1950–53 SALES, 33.3% ANNUAL RATE OF DEPRECIATION OF ADVERTISING EXPENDITURES ASSUMED (t-RATIO IN PARENTHESES)

Regression	Profit Rate	Intercept	Four-Firm Concentration Ratio (%)	Relative Market Share (%)	Advertising-to-Sales Ratio (%)	Change in Industry Demand (%)	Five-Digit-Firm Diversification Ratio (%)	Ratio of Advertising Investment to Net Worth (Z/E) (%)	$\frac{1}{\text{Log Assets}}$	R^2	Corrected R^2
Type (2)	1950–53 average true profit rate dependent										
3′	P_1^*	4.01	0.048 (1.75)	0.070 (2.73)	0.534 (1.69)	0.015 (1.05)	−0.038 (1.62)		0.952 (0.17)	.586	.510
4′	P_2^*	6.47	0.047 (1.68)	0.062 (2.39)	0.509 (1.59)	0.016 (1.08)	−0.047 (1.95)		−0.954 (0.17)	.573	.495
Type (3)	1950–53 average measured profit rate dependent										
5′	P_1	4.06	0.050 (1.73)	0.072 (2.63)	0.402 (0.96)	0.015 (1.01)	−0.040 (1.61)	0.441 (0.52)	0.296 (0.01)	.626	.545
6′	P_2	6.37	0.050 (1.69)	0.064 (2.30)	0.394 (0.95)	0.017 (1.12)	−0.047 (1.90)	0.241 (0.33)	−1.45 (0.25)	.610	.525

IV. Conclusions

Previous studies investigating the relationship between measured profit rates, advertising intensity, and market structure have consistently found a strong positive relationship between profitability and advertising intensity. Advertising intensity has generally been treated as a proxy for the degree of product differentiation, indicating that product differentiation is associated with high profitability. An alternative explanation for finding a strong positive relationship between measured profitability and advertising intensity is that advertising intensity is a proxy for errors in measured profit rates resulting from the expensing of advertising. The evidence presented above supports the latter explanation rather than the former. Although the evidence is based on only one relatively small sample of firms, it clearly indicates the need for caution in interpreting the results of regressions using measured profit rates as the dependent variable

The relationship between advertising and profitability has attracted attention from policymaking bodies such as the FTC because of the implications for a public policy aimed at maintaining the maximum degree of competition If advertising by differentiating products acts as a barrier to entry, a positive relationship between advertising and profitability can be expected The strong positive relationship between measured profit rates and advertising intensity found in previous studies of the relationship between advertising and profitability has lent support to the view that a corrective policy toward advertising is needed to maintain competition The evidence presented above indicates that a strong positive relationship between measured profit rates and advertising intensity disappears when the profit rates are adjusted to correct for the expensing of advertising Thus, the view that a positive relationship between measured profit rates and advertising intensity indicates the need for corrective policy toward advertising seems unwarranted.

It is interesting to note that, while adjusting measured profit rates to correct for the expensing of advertising casts doubt on the need for a corrective policy toward advertising to maintain competition, the need for a different kind of corrective policy toward advertising is clearly indicated The treatment of advertising expenditures as a current expense rather than as an investment in the calculation of reported profits reduces the reported profits of firms that make positive net investments in advertising The tax payments of such firms are reduced because the level of tax payments varies directly with the level of reported profits.[20] The data on net investments in advertising and marginal corporate income tax rates used in calculating estimated true profit rates for 40 food manufacturing firms imply that over the 4 years, 1950 through 1953, the 40 firms were able to avoid a total of $373 million of taxes due to the expensing of advertising for tax purposes[21] The average annual tax avoidance over the 4 years for each of the 40 firms is listed in table A1 The tax avoidance amounts to an average of more than $2 million per firm per year The ratio of tax avoidance to advertising expenditures is 0 46, which is the equivalent of a 46 percent subsidy on advertising expenditures[22]

---

[20] The reduction in current tax payments gives rise to a deferred tax liability which is incurred when the firm subsequently reduces its investment in advertising. For a firm which continually increases its advertising expenditures there is never a subsequent reduction in its advertising investment, so the deferred tax liability is never incurred.

Weiss (1969) proposes that to eliminate the implicit subsidy on advertising the capitalization of advertising expenditures for tax purposes should be enforced in cases where the advertising expenditures have long-lasting effects (pp. 429–30) The size of the implicit subsidy on the advertising of the 40 food manufacturing firms indicates the importance of the subsidy for firms that have large amounts of long-lasting advertising, and thus provides evidence in support of Weiss's proposal for tax reform

[21] The amount of tax avoidance was calculated by multiplying total net investment in advertising for the 40 firms in each year by the prevailing marginal corporate income tax rate and then summing over the 4 years For the marginal tax rates used, see n 16 above

[22] The amount of tax avoidance during the years 1950 through 1953 was to a degree uncharacteristically high due to the imposition of excess profits taxes after the beginning of the Korean war Assuming that the corporate income tax rate had been a normal 52 percent during the period and that the same level of advertising expenditures prevailed, the total tax avoidance for the firms would have been $249 million and the ratio of tax avoidance to advertising expenditures would have been 0.31 Thus, even with normal income tax rates the 40 food manufacturing firms would have been receiving the equivalent of a 31 percent subsidy on their advertising expenditures

TABLE A1

ROFIT RATES, ADVERTISING-TO-SALES RATIOS, ADVERTISING ASSETS, AND TAX AVOIDANCE FROM EXPENSING OF ADVERTISING, FOR 40 FOOD MANUFACTURING FIRMS

COMPANY	AVERAGE PROFIT RATES (%)				ADVERTISING-TO-SALES RATIO (%)	AVERAGE ADVERTISING ASSET ($ MILLIONS)	AVERAGE ANNUAL TAX AVOIDANCE ($ MILLIONS)
	P_1	P_2	P^*_1	P^*_2			
Swift & Co	6.5	7.2	5.8	6.3	0.37	63.6	4.32
Armour & Co	5.1	5.8	4.9	5.4	0.27	34.0	2.87
Schenley Industries	6.4	8.6	5.7	7.0	2.50	79.2	5.31
National Dairy Products	9.9	13.0	8.8	10.9	0.92	59.3	5.19
Jos Seagram & Sons	11.3	14.2	9.1	10.5	2.06	101.6	6.92
General Foods	10.7	12.5	7.2	7.7	4.03	182.0	13.34
National Distillers Prod	6.4	8.0	5.8	6.9	1.96	59.6	4.40
Borden Co	9.0	10.9	8.1	9.4	0.60	33.4	1.98
Coca-Cola Co	16.6	16.8	12.5	12.6	2.23	60.4	1.50
National Biscuit Co	12.3	12.5	10.7	10.8	1.28	29.6	1.99
Hiram Walker & Sons	13.7	14.0	11.3	11.4	1.14	37.2	1.32
Corn Products Refining	12.4	12.4	11.0	11.0	1.32	21.2	1.01
California Packing	8.1	9.2	7.5	8.2	1.13	16.6	1.18
Wilson & Co	4.3	4.4	4.1	4.2	0.11	8.5	0.30
General Mills	9.4	9.7	6.0	6.1	3.24	107.0	7.17
Publicker Industries	4.1	4.1	4.1	4.1	1.57	12.3	1.41
H J Heinz	6.5	6.7	4.9	4.8	1.45	41.1	0.66
Standard Brands	8.3	8.4	5.3	5.3	1.94	80.8	2.09
Libby, McNeil, & Libby	6.5	7.8	5.2	5.8	1.11	24.0	0.70
Wesson Oil & Snowdrift	7.0	7.0	6.5	6.5	1.02	11.4	0.93
Pillsbury Mills	6.1	7.1	5.4	5.9	2.49	34.8	4.30
Pabst Brewing	11.1	12.1	9.6	10.3	1.75	17.2	1.45
Quaker Oats Co	11.3	13.0	7.9	8.4	3.23	50.2	3.52
Ralston Purina Co	10.6	12.3	8.4	9.3	0.63	27.6	0.93
Carnation Co	10.5	12.4	9.0	10.1	1.01	18.8	1.52
Wm Wrigley, Jr Co	15.8	15.8	11.4	11.4	5.73	30.5	2.05
Continental Baking Co	9.8	12.3	7.6	8.9	1.16	19.4	0.77
Stokely–Van Camp	8.1	8.1	7.1	7.0	1.26	9.6	0.73
Sunshine Biscuits	15.2	15.2	13.0	13.0	1.07	8.5	0.56
Brown-Forman Distillers	9.3	12.4	8.0	9.6	3.68	12.8	1.40
Kellogg Co	23.5	23.5	10.7	10.7	5.04	55.8	3.08
Pet Milk Co	7.8	7.8	6.4	6.4	1.38	13.4	1.14
Best Foods	13.8	13.8	9.4	9.4	2.85	21.2	0.91
Geo Hormel & Co	8.7	10.0	7.3	8.0	0.46	9.5	0.62
Canada Dry Ginger Ale	8.5	9.6	6.1	6.4	3.70	18.4	1.01
Glenmore Distillers Co.	7.9	9.7	6.4	7.1	3.89	14.8	1.17
Hunt Foods	9.5	9.9	8.8	9.0	3.64	7.1	1.45
Pepsi-Cola Co	11.4	12.4	8.3	8.6	4.13	16.1	0.92
Gerber Products Co	15.1	16.9	12.0	12.7	2.32	7.4	0.75
Green Giant Co	8.6	10.2	7.1	7.7	3.61	8.7	0.78

References

Bloch, Harry. "Advertising, Competition, and Market Performance." Ph.D thesis, Dept Econ, Univ. Chicago, 1971.

Comanor, William S, and Wilson, Thomas A "Advertising, Market Structure, and Performance." *Rev. Econ. Statis.* 49, no 4 (November 1967): 423–40.

Federal Trade Commission. "Economic Report on the Influence of Market Structure on the Profit Performance of Food Manufacturing Companies" Washington. Government Printing Office, 1969.

Goldberger, Arthur S *Econometric Theory* New York Wiley, 1964.

Imel, Blake, and Helmberger, Peter "Estimation of Structure-Profit Relationships with Application to the Food Processing Sector." *A E R* 61, *no* 4 (September 1971) 614–27

Johnston, J *Econometric Methods.* New York McGraw-Hill, 1963

Lambin, Jean-Jacques "Optimal Allocation of Competitive Marketing Efforts An Empirical Study" *J Bus* 43, *no* 4 (October 1970): 468–84

Miller, Richard A "Market Structure and Industrial Performance Relation of Profit Rates to Concentration, Advertising Intensity, and Diversity" *J Indus Econ* 27, no. 2 (April 1969)· 104–18

Nerlove, Marc, and Waugh, Frederick V "Advertising without Supply Control" *J Farm Econ* 42, no. 4 (November 1961): 813–37

Palda, Kristain *The Measurement of Cumulative Advertising Effects* Englewood Cliffs, N J Prentice-Hall, 1964

Peles, Yoram "Rates of Amortization of Advertising Expenditure." *J.P.E* 79, no 5 (September–October 1971) 1032–58

Shepherd, William G. "The Elements of Market Structure" *Rev. Econ Statis* 54, no 1 (February 1972) 25–37

Telser, Lester G. "Advertising and Cigarettes" *J.P E* 70, no 5 (October 1962) 471–99

——— "Some Aspects of the Economics of Advertising" *J Bus* 41, no. 2 (April 1968) 166–73

——— "Theory of the Firm—Discussion" *A E R* 59, no 2 (May 1969)· 121–23

Theil, Henri *Principles of Econometrics.* New York Wiley, 1971.

Weiss, Leonard W "Advertising, Profits, and Corporate Taxes" *Rev Econ Statis* 51, no 4 (November 1969) 421–30

Investment Incentives and the Corporate Income Tax

Agnar Sandmo

Norwegian School of Economics and Business Administration

Within the framework of the neoclassical theory of investment and capital this article discusses the effect of the corporate income tax on investment incentives, particularly with respect to capital goods of different degrees of durability Sources of distortion are sought in the treatment of depreciation, in incomplete interest deductions, and in the treatment of capital gains Depending on the relative magnitudes of the parameters involved, the corporate income tax may change relative prices in favor of either short-term or long-term capital goods

1. Introduction

The purpose of this article is to present an analysis of the effect of the corporate income tax on investment incentives The framework of analysis is the neoclassical theory of investment and capital, the firm's optimal use of labor and capital over time is derived from the basic criterion of present-value maximization A large part of the modern work in this area has been done by Dale Jorgenson and his associates, they have also considered the effect of the corporate income tax, for example, in the article by Hall and Jorgenson (1967), and subjected the model to careful empirical tests [1] There still seems to be a need, however, for a more detailed theoretical discussion of the model, particularly since there is a competing body of theory—exemplified by writings of E. Cary Brown (1948), Leif Johansen (1965), and Jan Södersten (1972)—which appears to lead to somewhat different con-

---

I am indebted to Leif Johansen, Thore Johnsen, and the referees for extremely helpful comments and suggestions.

[1] In addition to the Hall-Jorgenson article, the reader is referred to the book by Fromm (1971) for a series of interesting, although perhaps rather inconclusive, empirical investigations of the problem.

clusions and which also raises some new questions. Briefly, these writers have turned their attention to the effect of the corporate income tax on the composition of the capital stock, in particular with respect to durability, while the neoclassical analysis has been focused on the substitution between labor and capital. The present paper attempts to integrate the two approaches by extending the neoclassical model to take account of the existence of capital goods of different degrees of durability

The analysis is partial equilibrium in character, although some general-equilibrium considerations are touched upon in Section 9 A full-scale investigation of the general-equilibrium implications of the analysis would be of considerable interest, but it lies beyond the scope of a single paper

2. The Basic Model of Capital Accumulation

The formulation of the model of capital accumulation to be employed in the following is derived from an earlier paper (Sandmo 1971*a*); the only change of any significance is that we shall work with two different capital goods instead of one [2] In this section we shall present the basic features of the model while postponing the analysis of taxation to Section 3

Output in period t depends on the quantities of labor and the two types of capital used during the period

$$X_t = F(L_t, K_{1t}, K_{2t}) \tag{1}$$

Marginal productivities are assumed to be everywhere positive, and there are nonincreasing returns to scale

Gross investments in period t, I_{jt} ($j = 1, 2$), are equal to net investments plus replacements We take depreciation to be proportional to capital stock, so that we have, for each capital good,

$$I_{jt} = K_{j,t+1} - K_{jt} + \delta_j K_{jt} \qquad (j = 1, 2, 0 < \delta_j < 1) \tag{2}$$

It is also assumed that the stocks of capital at the beginning of the initial period are given, that is,

$$K_{j0} = \text{exogeneously given} \qquad (j = 1, 2) \tag{3}$$

These assumptions imply that capital goods bought in period t increase the productive capacity of the firm only in period $t + 1$ We may think of this lag as the time required to install new capital goods. It is, of course, a weakness of the model that the length of this time period is taken to be exogeneously determined, so that adjustment costs are ignored.

---

[2] As in the earlier paper, the model is formulated in discrete time An alternative would have been to use continuous time and to derive the results by means of calculus-of-variations methods Such a reformulation presents no essential problems and would in some respects have been more elegant. But the problems discussed in this paper can be handled equally well by the simpler approach chosen here, which probably appeals to a wider readership.

The firm is assumed to be a price taker in both commodity and factor markets In period t the price of output is p_t, w_t is the wage rate, and q_{1t} and q_{2t} are the prices of capital goods For simplicity we shall assume that these prices are expected to be constant over time, so that $p_t = p$, $w_t = w$, and $q_{jt} = q_j$ ($j = 1, 2$) Likewise, the rate of interest is constant over time and equal to i, it is assumed to be strictly positive Some special problems associated with price variability are discussed in Section 7

The net cash flow of the firm in period t—the excess of receipts over expenditure—is defined as

$$R_t = pX_t - wL_t - q_1I_{1t} - q_2I_{2t} \tag{4}$$

The objective of the firm is to maximize the present value of its cash flow, that is,

$$V = \sum_{t=0}^{\infty} (1 + i)^{-t}R_t \tag{5}$$

By substituting into this expression from (1), (2), and (4), we can write the maximand more fully as

$$\begin{aligned} V = \sum_{t=0}^{\infty} (1 + i)^{-t}[&pF(L_t, K_{1t}, K_{2t}) - wL_t \\ &- q_1(K_{1,t+1} - K_{1t} + \delta_1K_{1t}) \\ &- q_2(K_{2,t+1} - K_{2t} + \delta_2K_{2t})] \end{aligned} \tag{6}$$

The first-order maximum conditions then become

$$p\frac{\partial F}{\partial L_t} - w = 0 \qquad (t = 0, 1, 2, \quad) \tag{7}$$

$$p\frac{\partial F}{\partial K_{1t}} - (i + \delta_1)q_1 = 0 \qquad (t = 1, 2, \quad) \tag{8}$$

$$p\frac{\partial F}{\partial K_{2t}} - (i + \delta_2)q_2 = 0 \qquad (t = 1, 2, \quad) \tag{9}$$

Note that the t-index runs from $t = 0$ for labor and from $t = 1$ for capital goods. This is because the initial capital stocks are given, so that the only production decision in the initial period concerns the amount of labor to be employed with the given amounts of real capital.

Equations (7)–(9) state the equality between the value of each factor's marginal product and its price In the case of labor the interpretation of the condition is straightforward. In the case of capital goods, however, the factor prices involved are the shadow prices or rental rates, which represent the cost of holding a unit of capital goods for one time period. This is obviously equal to the interest on the value of a unit of capital plus the value of its depreciation over the period.

Since we have now derived the shadow prices of capital goods, we can define profit in period t as

$$\pi_t = pF(L_t, K_{1t}, K_{2t}) - wL_t - (i + \delta_1)q_1K_{1t} - (i + \delta_2)q_2K_{2t}, \tag{10}$$

and the present value of profits as

$$V' = pF(L_0, K_{10}, K_{20}) - wL_0 + \sum_{t=1}^{\infty} (1 + i)^{-t}\pi_t \tag{11}$$

Obviously, maximization of V' with respect to L_t and K_{jt} will yield the first-order conditions (7)–(9) Accordingly, we have shown that maximization of the present value of cash flows as defined by (4) and maximization of the present value of profits as defined by (10) lead to identical production and investment decisions, so that they are equivalent decision rules

As has been shown in my earlier paper (1971*a*) and as will emerge from the discussion in Section 7 below, the simple optimality criterion of equating marginal products and factor prices in each period is a special case of a myopic decision rule which can be generalized to the case of prices that are changing over time, in that case, only the current expected rates of price change enter into the rental rates of capital goods This property of the optimal policy greatly simplifies the analysis, but it does rest on some rather restrictive assumptions It assumes the existence of capital markets that are perfect in the extreme sense that capital goods can be bought and sold in any amount at the same price It also assumes that adjustment costs are either insignificant or that they are simply proportional to the amount of gross investment, so that these costs can be interpreted as included in capital goods prices, q_j

These are strong assumptions, which clearly limit the applicability of the analysis The first one, however, may not be as restrictive as it seems, since the analysis will still be valid, even if gross investment is restricted to be nonnegative provided that the amount of net disinvestment does not exceed the amount of depreciation The second assumption is more serious in that it may limit the possibilities of providing hypotheses about the response of investment demand to tax changes in the very short run Still, in the present context the more long-run effects of taxes on rental rates of capital goods seem to be the more interesting ones

Returning to our present-value decision rules, it is well known that these rules can themselves be derived from a more fundamental criterion, namely, utility maximization over time on the part of the owners of the firm. If owners can borrow and lend in a perfect capital market, the policy which from their point of view is the optimal one for the firm is to ensure that they attain the highest possible budget constraint for intertemporal transactions [3] This policy is precisely the maximization of the present value of cash flows or profits

3. Neutrality of the Corporate Income Tax

In order to discuss the effects of a tax on profits we must ensure that the model implies the existence of profits In long-run competitive equilibrium profits will be zero, and the corporate income tax will not yield any revenue In order to have a positive tax base it seems as if we shall have to depart from the competitive assumptions set out above This can be done without much difficulty, it is, in fact, easy to show that the first-order conditions for a firm which is a monopolist in the product market differ from (7)–(9) only in having marginal revenue substituted for price Thus, the following analysis can be given a more general interpretation by simply interpreting p as marginal revenue

However, even in the perfectly competitive case with constant returns to scale, it may be meaningful to analyze the corporate income tax if taxable profit is not equal to "true" profit If some of the imputed factor rewards are not in fact tax deductible, a tax on profit will really amount to a tax on those factors We shall return to several examples of this later on

Let us first suppose that a proportional profits tax is added to the model described in the previous section In what way will the imposition of such a tax influence the allocation of capital and labor over time? One important conclusion is immediate if taxable profit is defined as in (10), a tax on profit will be equivalent to a tax on present value and will have no effects on investment decisions [4]

Analytically this result is, of course, quite trivial But it is worth stressing that the introduction of the time dimension into the model of the firm does not necessarily imply that the corporate income tax loses its neutrality, neutrality can be achieved by proper definition of the tax base

4. The Role of Depreciation Allowances

The imposition of a corporate income tax usually implies a set of rules governing the depreciation allowances that can be deducted from gross income It seems to be quite generally believed that it is in the interaction between the depreciation allowances and the tax rate that one finds the most important distortive effects of the corporate income tax.

---

[3] This point is elaborated in Hirshleifer (1958) and Sandmo (1968), who also consider the effect of imperfections in the capital market

[4] In somewhat different versions this conclusion can also be found in the articles by Samuelson (1964) and Hall and Jorgenson (1967).

Profit before tax is still defined as in (10), while taxable profit is now

$$\pi_t^* = pF(L_t, K_{1t}, K_{2t}) - wL_t - (i + \alpha_1)q_1K_{1t} - (i + \alpha_2)q_2K_{2t}, \quad (12)$$

where α_j are the rates of depreciation allowance. If s is the tax rate, the optimal policy is found by maximizing, in each current period, after-tax profit which is equal to $\pi_t - s\pi_t^*$. The first-order conditions then become

$$p\frac{\partial F}{\partial L_t} - w = 0, \qquad (t = 0, 1, 2, \ldots) \quad (13)$$

$$p\frac{\partial F}{\partial K_{1t}} - \left[i + \delta_1 + \frac{s}{1-s}(\delta_1 - \alpha_1)\right]q_1 = 0, \quad (14)$$
$$(t = 1, 2, \ldots)$$

$$p\frac{\partial F}{\partial K_{2t}} - \left[i + \delta_2 + \frac{s}{1-s}(\delta_2 - \alpha_2)\right]q_2 = 0 \quad (15)$$
$$(t = 1, 2, \ldots)$$

We see that when the tax depreciation allowances do not correspond to the "true" rates of depreciation, the tax rate does influence the shadow prices of capital goods and thereby the choice of optimal factor proportions.[5]

There are two ways in which the corporate income tax may lead to a distortion of pretax factor prices. First, it may increase rental rates relative to the wage rate and thus lead to a less capital-intensive choice of technique. Second, it may influence the relative rental rates. Of course, both effects may be present simultaneously.

The first effect will be unambiguous in the cases where depreciation allowances are either both greater than or both less than the true rates of depreciation; these cases create incentives to substitute capital for labor and labor for capital, respectively.[6] The ambiguous case is the intermediate one where one capital good is taxed and the other one is subsidized relative to labor. In this case the corporate income tax has no obvious effect on the general capital intensity of production.

To study the effects on the composition of the capital stock, we compare the rental rates of the two types of capital before and after tax. Thus, we define:

---

[5] We do not carry through here a full comparative-static analysis. The interesting part of the theory concerns the influence of taxation on the relative shadow prices of factors of production. The further analysis of the effects of factor price changes on factor demands is a standard exercise in the theory of the firm.

[6] It is theoretically possible, although practically implausible, that rental rates might become negative. This case is meaningless in the context of a perfect capital market and is not taken care of by the present formulation.

$$r_j^0 = (i + \delta_j)q_j, \qquad (j = 1, 2) \tag{16}$$

$$r_j^s = \left[i + \delta_j + \frac{s}{1 - s}(\delta_j - \alpha_j)\right] q_j \qquad (j = 1, 2) \tag{17}$$

We shall say that the corporate income tax discriminates in favor of capital of type 1, is neutral with respect to capital-goods inputs, or discriminates in favor of capital of type 2 according to whether the relative rate of increase in the rental rate of type 1 is less than, equal to, or greater than that of type 2, that is,

$$\frac{r_1^s}{r_1^0} \lesseqgtr \frac{r_2^s}{r_2^0} \tag{18}$$

Substituting from (16) and (17) and rearranging terms, we can write this condition as

$$\frac{\delta_1 - \alpha_1}{i + \delta_1} \lesseqgtr \frac{\delta_2 - \alpha_2}{i + \delta_2} \tag{19}$$

This is the general condition for the discriminatory effect as between capital goods of the corporate income tax For the tax to be neutral with respect to the composition of capital, the gaps between the rates of true depreciation and depreciation allowance must constitute the same fraction of the rate of true depreciation plus interest

There are two special cases in which condition (19) is simplified and which seem to be of considerable practical significance The first case is the one where the rate of depreciation allowance is a flat one, that is, $\alpha_1 = \alpha_2 = \alpha$ Condition (19) is then reduced to

$$\delta_1 \lesseqgtr \delta_2 \tag{20}$$

But this means simply that in the case of a flat rate of depreciation allowance, the corporate income tax creates an incentive to change the composition of the capital stock in favor of more durable capital goods

Condition (19) can be rewritten as

$$\frac{\delta_1 - \alpha_1}{\delta_1} \frac{\delta_1}{i + \delta_1} \lesseqgtr \frac{\delta_2 - \alpha_2}{\delta_2} \frac{\delta_2}{i + \delta_2} \tag{21}$$

Suppose now that the tax authorities fix the rates of depreciation allowance such that they are proportional to true rates of depreciation, that is,

$$\frac{\delta_1 - \alpha_1}{\delta_1} = \frac{\delta_2 - \alpha_2}{\delta_2} \tag{22}$$

In the case where the numerators of the fractions are positive (the case of decelerated depreciation), we get again condition (20) We conclude that in the case of decelerated depreciation and proportional depreciation allowances, the corporate income tax changes investment incentives in favor of more durable capital goods When the numerators are negative, however, substitution of (22) into (21) will yield condition (20) with the inequality signs reversed Consequently, in the case of accelerated depreciation and proportional depreciation allowances, the corporate income tax changes investment incentives in favor of less durable capital goods

The intuitive reason for these results should be clear The rental rates have two components, interest and depreciation In the case of decelerated depreciation the corporate income tax is—in addition to being simply a tax on profit—a tax on depreciation If depreciation is already the major component of the rental rate, such a tax is more important than if depreciation is initially a very insignificant part of the opportunity cost of capital, that is, the tax falls more heavily on short-term than on long-term capital goods In the case of accelerated depreciation, the tax contains an element of a capital subsidy, and the argument applies in reverse

5. Comparison with the Brown-Johansen Formulation

An influential early contribution to the analysis of the tax treatment of depreciation is that of Brown (1948), his analysis was later extended and modified by Musgrave (1959), Johansen (1965), and Sodersten (1972) It is obviously of interest to compare their formulation with the present one A direct comparison is difficult, however, the authors cited do not derive their conclusions from an explicit optimization model for the firm, but rather from simple present-value calculations for a single investment project Moreover, there is the additional difficulty that they do not assume a constant rate of depreciation Still, a comparison, if only in more general terms, seems to be called for The following exposition is based on that of Johansen (1965, chap 7, par 5)

Johansen writes the present value of an investment project before tax as

$$V = a \sum_{t=1}^{n} (1 + i)^{-t} - K = a\beta_n - K \tag{23}$$

Here a is the cash flow resulting from investing the amount K in the initial period, this cash flow is assumed to be constant over time The tax base in year t ($t = 1, 2, \quad , n$) is defined as net income, which is $a - (1/n)K$, where n is the lifetime of the project and a linear depreciation allowance is assumed Present value after tax is then equal to present value before tax minus the present value of tax payments, that is,

$$V^* = V - s\left(a - \frac{K}{n}\right)\beta_n. \tag{24}$$

For $t > 0$ we must have $\beta_n < n$ Hence, for a project that was profitable before tax, so that $V > 0$, we must have $a - K/n > 0$ Equation (24) then implies that a project with positive present value before tax may have negative present value after tax, thus, the corporate income tax is not neutral with respect to the choice between capital and other factors of production. Moreover, since the present value of tax payments is an increasing function of n, it follows that the tax will also discriminate against long-term investment projects

What is the assumption made here about true depreciation? Since a is constant over the finite lifetime of the investment, it must be assumed that depreciation is of the "sudden death" variety, the capital stock is constant until the end of the nth period, when it depreciates 100 percent But this means that we have a special case of accelerated depreciation allowances, and we showed in the previous section that this system tends to discriminate against investment in more durable capital goods Thus, the two analyses are consistent on this point

This comparison may not go to the heart of the matter, however Evaluating an investment on the basis of a fixed lifetime does not go very well with the assumption of perfect reversibility, for in the latter case a myopic policy would be optimal An alternative interpretation of the Brown-Johansen formulation is therefore that capital is irreversible, once installed, it cannot be resold The implication of this is clearly that in value terms the rate of depreciation is 100 percent in the initial period, although the technical lifetime of the capital good is still n periods With these assumptions it becomes very hazardous to compare the two approaches, for the one adopted in the present paper cannot be easily extended to cover the case of irreversible investments There is a need, therefore, to construct models of optimal capital policy that incorporate the Brown-Johansen formulations in a more comprehensive framework

6. Equity Capital and Incomplete Interest Deductions

Tax laws in most countries do not provide for deductibility of interest on the full value of the capital stock The nondeductibility of interest on equity capital has been seen by many writers as the main reason why the corporate income tax cannot be assumed to be allocatively neutral

The present model does not lend itself very well to an analysis of this problem A really satisfactory analysis would have to involve uncertainty and the determination of the optimal debt-equity ratio via the price formation in capital markets. These are complex and unsettled questions and will not be explored here. Instead, we shall take as a given fact that part of the firm's capital stock is financed by equity and study the consequences of this on the incentives to hold capital goods Suppose first that the firm's equity capital is tied up with a specific capital stock which is not subject to change, for example, the land used by the firm or buildings which are specific to the firm and indivisible Otherwise, all other capital is debt financed, so that interest is deductible. It is easy to see that in this case the corporate income tax must be allocatively neutral, since the required debt-equity ratio is irrelevant to marginal decisions. The interesting case is where there is a required debt-equity ratio for capital goods purchases that are determined by the firm's optimal policy.

Assuming now that depreciation allowances correspond to true depreciation, taxable profit is equal to

$$\begin{aligned}\pi_t^* &= pF(L_t, K_{1t}, K_{2t}) - wL_t \\ &\quad - \delta_1 q_1 K_{1t} - \delta_2 q_2 K_{2t} - i(1 - k)(q_1 K_{1t} + q_2 K_{2t}),\end{aligned} \tag{25}$$

where k is the required ratio of equity to the total value of the firm's capital stock ($0 < k < 1$) Maximization of net profit, $\pi - s\pi^*$, yields the first-order conditions

$$p\frac{\partial F}{\partial L_t} - w = 0, \qquad (t = 0, 1, 2, \quad) \tag{26}$$

$$p\frac{\partial F}{\partial K_{1t}} - \left[i\left(1 + \frac{s}{1 - s}k\right) + \delta_1\right]q_1 = 0, \qquad (t = 1, 2, \quad) \tag{27}$$

$$p\frac{\partial F}{\partial K_{2t}} - \left[i\left(1 + \frac{s}{1 - s}k\right) + \delta_2\right]q_2 = 0 \qquad (t = 1, 2, \quad) \tag{28}$$

A positive value of k increases the rental rates relative to the wage rate and creates an incentive to substitute labor for capital This effect is rather obvious It should also be noted, however, that the relative rental rates are also changed, given that they have different rates of depreciation The equity requirement will change relative prices in favor of capital of type 1 or type 2 according to whether

$$\frac{i\{1 + [s/(1 - s)]k\} + \delta_1}{i + \delta_1} \lesseqgtr \frac{i\{1 + [s/(1 - s)]k\} + \delta_2}{i + \delta_2} \tag{29}$$

This condition simplifies to

$$\delta_1 \gtreqless \delta_2 \tag{30}$$

The conclusion is, then, that incomplete interest deductibility changes the relative rental rates in favor of short-term capital goods.

This effect, though usually neglected in discussions of the corporate income tax, is hardly very surprising As we have seen, incomplete deductibility raises the effective rate of interest A classic result of capital theory is that a higher rate of interest leads to a substitution of short-term for long-term capital goods, and it is precisely this effect that we observe here

7. Distortions Induced by Price Changes

It is, of course, a simplification to work with a model in which prices are assumed to be constant over time, in fact, price variability may be one of the more important reasons for differences between "true" and taxable profit. If p_t, q_t, and w_t are no longer constant with respect to t, it is easy to see that profit before tax must be equal to

$$\begin{aligned}\pi_t = {} & p_t F(L_t, K_{1t}, K_{2t}) - w_t L_t \\ & - [i + \delta_1 - (1 + i)\gamma_{1t}] q_{1t} K_{1t} \\ & - [i + \delta_2 - (1 + i)\gamma_{2t}] q_{2t} K_{2t},\end{aligned} \tag{31}$$

where we have defined the rates of change of capital goods prices as

$$\gamma_{jt} = \frac{q_{jt} - q_{j,t-1}}{q_{jt}} \qquad (j = 1, 2) \tag{32}$$

The last terms in the bracketed expressions are the capital-gains components of current profit[7] For the corporate income tax to be completely neutral, it is necessary that these components are also included in taxable profit; gains must be taxed and losses must be deductible If this is not the case, the existence of the corporate income tax will raise the rental rates relative to the wage rate if prices of capital goods are falling and reduce the rental rates in the opposite case

Failure to allow for capital gains may also affect the composition of the capital stock If taxable profit is written as

$$\pi_t^* = p_t F(L_t, K_{1t}, K_{2t}) - w_t L_t - (i + \delta_1) q_{1t} K_{1t} - (i + \delta_2) q_{2t} K_{2t}, \tag{33}$$

so that the firm maximizes $\pi_t - s\pi_t^*$, the first-order conditions are

$$p_t \frac{\partial F}{\partial L_t} - w_t = 0, \qquad (t = 0, 1, 2, \quad) \tag{34}$$

$$p_t \frac{\partial F}{\partial K_{1t}} - \left[i + \delta_1 - \frac{1}{1 - s}(1 + i)\gamma_{1t}\right] q_{1t} = 0, \qquad (t = 1, 2, \quad) \tag{35}$$

$$p_t \frac{\partial F}{\partial K_{2t}} - \left[i + \delta_2 - \frac{1}{1 - s}(1 + i)\gamma_{2t}\right] q_{2t} = 0 \qquad (t = 1, 2, \quad) \tag{36}$$

Substituting from (35) and (36) into (18), we see that the tax will change prices in favor of capital of type 1, lead to no change, or to a change in favor of capital of type 2 according to whether

---

[7] See n 6 above

$$\frac{i + \delta_1 - [1/(1-s)](1+i)\gamma_{1t}}{i + \delta_1 - (1+i)\gamma_{1t}} \lesseqgtr \frac{i + \delta_2 - [1/(1-s)](1+i)\gamma_{2t}}{i + \delta_2 - (1+i)\gamma_{2t}} \tag{37}$$

This condition can be simplified to read

$$(i + \delta_1)\gamma_{2t} \lesseqgtr (i + \delta_2)\gamma_{1t}. \tag{38}$$

As was to be expected, the result depends crucially on the relative rates of change of capital-goods prices—hardly a very interesting observation. What may not be so obvious, however, is that an effect remains even when the relative rates of change of the two capital goods prices are the same, so that, $\gamma_{1t} = \gamma_{2t}$. If this common rate of change is positive, (38) becomes

$$\delta_1 \lesseqgtr \delta_2 \tag{39}$$

If capital goods prices increase at a uniform rate, the failure to tax capital gains creates an incentive to substitute long-term for short-term capital goods.

If the common rate of change is negative, condition (38) becomes

$$\delta_1 \gtreqless \delta_2 \tag{40}$$

If capital goods prices decrease at a uniform rate, failure to allow tax deductions for capital losses creates an incentive to substitute short-term for long-term capital goods.

It may be intuitively helpful to think of a uniform increase of capital-goods prices as a lowering of the real rate of interest, we know that this favors long-term investment. Similarly, decreasing prices of capital goods means a higher real rate of interest and a higher return on short-term investments. Note, however, that the conclusions are independent of other rates of price change and therefore are much more general than statements about the impact of general inflation or deflation.

8. Some Neglected Problems: Imperfect Capital Markets and Uncertainty

The foregoing analysis depends crucially on the assumption of a perfect capital market. Not only is the firm facing a given market rate of interest which is insensitive to the amount it borrows, the availability of credit is also independent of the firm's own income and capital position. The latter assumption may be of particular importance in this connection, since the opinion is widely held that the effect of the corporate income tax on investment decisions works mainly through its reduction of internal financial funds.

Models incorporating such ideas are not easy to formulate They would clearly imply that one must go beyond the framework of the present analysis to study an optimizing model of the supply side of the credit market as well as an equilibrium theory explaining the market allocation of funds between equity and debt These considerations raise unsettled theoretical problems of great complexity which we cannot attempt to tackle here Apart from this, however, the significance of these complications may easily be overstated. Seen from the viewpoint of a single firm, such "income effects" may seem highly important, but that is frequently because the single firm does not take into account that the same effects are present for all firms After market prices have adjusted to the tax, the availability of internal funds may not be a significant independent force acting on the demand for capital goods

No mention has been made so far of the effects of uncertainty This opens up a whole new set of problems, particularly because uncertainty may enter in many ways, and several alternative assumptions may be made about the objectives of the firm in an uncertain environment Perhaps the simplest assumption would be that the firm maximizes expected present value, one might then believe that the analysis would remain valid with a simple reinterpretation of the variables in terms of expected values But it is in fact obvious that this will be true only under some additional assumptions, if, for example, the price of the product and the technology are both uncertain and stochastically dependent, the equivalence will not hold, simply because the expectation of the product of price and output will not be equal to the product of the expectations In general, the reinterpretation of the model will be valid only if the conditions of the Simon-Theil certainty equivalence theorem are satisfied [8] Although these are restrictive conditions, it may perhaps not be unreasonable to see our analysis as an approximation to the case of expected present-value maximization

A further complication enters if it is assumed that the firm is risk averse Not only will we then face the problem of deriving meaningful shadow prices, there is the added problem that the influence of factor prices on optimal factor use becomes much more complex, since uncertainty introduces income effects into the comparative statics of firm behavior.[9]

In one important respect the introduction of uncertainty may add significant new elements to the analysis As shown by Mossin (1968) and others,[10] proportional income taxation will affect the investor's choice between safe and risky assets In the case of full loss offset provisions, taxation will lead to a substitution of risky for safe assets in investors' wealth portfolios Suppose now that short-term capital goods are more risky than long-term ones, the precise meaning of this assumption would of course have to be clarified in a rigorous analysis. With insufficient depreciation allowances the corporate income tax will tend to discourage short-term investment according to the analysis in Section 4 above, but the risk aspects might tend to counteract this tendency This is, of course, only one of many possible outcomes and is not intended as more than an example, there is clearly a need, however, for a more detailed study of the interaction of durability and riskiness in the impact of the corporate income tax

---

[8] For an exposition of this theorem, see Theil (1964, chap. 2)

[9] Analysis of the behavior of the risk-averse firm may be found in Sandmo (1971*b*) and Leland (1972)

[10] The first treatment of this problem was by Domar and Musgrave (1944). A later discussion along the same lines can be found in Musgrave (1959, chap. 14)

9. Some Concluding Remarks

The above analysis is in the nature of partial-equilibrium theory We have studied the impact of the corporate income tax on the demand functions for capital and labor services An extension of the analysis would be to construct a general equilibrium model and to study the incidence of the tax, that is, its effects on market prices and the interest rate This problem has been studied by Harberger (1962), and the more recent literature has been surveyed by Mieszkowski (1969) It lies outside the scope of the present article to tackle these problems as well, but some brief remarks may be in order

It may be argued that a partial-equilibrium treatment like the present one tends inevitably to be misleading Suppose that one concludes that imposition of a corporate income tax means raising the rental rates of capital-goods services relative to the wage rate and taxing the use of long-term capital goods more heavily than the use of short-term capital goods Will this not tend to change general-equilibrium prices in such a way as to counteract these first effects? It has been maintained, for example, that the rate of interest will be lower in the new equilibrium situation, this will of course lower the rental rates generally and relatively more for long-term than for short-term capital goods

It is quite clear that taking account of the full general equilibrium effects of the tax will tend to modify the quantitative magnitudes of the effects analyzed above However, since a theoretical treatment can yield only qualitative results, this is hardly a very significant conclusion. The real question is whether the counteracting tendencies could actually be expected to be stronger than the primary effects of the tax; this does in fact seem very doubtful. For a general-equilibrium model to yield such results, it would probably have to violate even very weak stability conditions, which would tend to reduce further the plausibility of such a result.

We have seen that there are a variety of reasons why the corporate income tax may distort investment decisions, even on strict neoclassical assumptions If the chief rationale for the existence of the tax is to be sought in its redistributional effects, these distortions must be judged to be undesirable side effects, to be avoided if possible But a concept of taxable profits which correctly takes account of depreciation, interest, and capital gains may present the tax authorities with great administrative difficulties An alternative solution, which has been suggested by Vernon Smith (1963), would be to use cash flow as the tax base instead of profit, so that all new capital-goods purchases are written off for tax purposes immediately, while no capital costs are deductible in later periods Going back to equations (4) and (5), one sees immediately that this amounts to a tax on present value, which will be allocatively neutral, not only among capital goods of different degrees of durability, but also between labor and capital goods Of course, in years of large capital-goods purchases, cash flow may become negative, requiring either a subsidy or provisions for carrying forward the "deficit" to years of positive cash flows However, it is important to note that no general reduction of the tax base is involved, compared to the case of an "ideal," nondistortive definition of taxable profit From a theoretical point of view, this proposal throws an interesting light on the distortion arising from corporate income taxation, and its practical merits would seem to deserve serious attention

References

Brown, E Cary "Business-Income Taxation and Investment Incentives" In *Income, Employment, and Public Policy Essays in Honor of Alvin H Hansen* New York Norton, 1948 Reprinted in *Readings in the Economics of Taxation*, edited by R A Musgrave and C S Shoup London· Allen & Unwin, 1959

Domar, Evsey D , and Musgrave, Richard A "Proportional Income Taxation and Risk-Taking" *Q J E.* 58 (May 1944) 388–422

Fromm, Gary, ed *Tax Incentives and Capital Spending* Amsterdam North-Holland, 1971.

Hall, Robert E , and Jorgenson, Dale W "Tax Policy and Investment Behavior." *A E R* 57 (June 1967) 391–414

Harberger, Arnold C "The Incidence of the Corporation Income Tax" *J P.E* 70 (June 1962) 215–40

Hirshleifer, Jack "On the Theory of Optimal Investment Decision" *J P E* 66 (August 1958). 329–52

Johansen, Leif *Public Economics* Amsterdam North-Holland, 1965.

Leland, Hayne E "Theory of the Firm Facing Uncertain Demand" *A E.R.* 62 (June 1972) 278–91.

Mieszkowski, Peter. "Tax Incidence Theory. The Effect of Taxes on the Distribution of Income." *J Econ. Literature* 7 (December 1969) 1103–24

Mossin, Jan "Taxation and Risk-Taking An Expected Utility Approach." *Economica* 35 (February 1968). 74–82.

Musgrave, Richard A. *The Theory of Public Finance*. New York: McGraw-Hill, 1959.

Samuelson, Paul A. "Tax Deductibility of Economic Depreciation to Ensure Invariant Valuations." *J.P.E.* 72 (December 1964): 604–6.

Sandmo, Agnar. "Sparing og investering under sikkerhet og usikkerhet." *Statsøkonomisk Tidsskrift* 82 (September 1968): 143–63.

———. "Investment and the Rate of Interest." *J.P.E.* 79 (November/December 1971): 1335–45 (*a*)

———. "On the Theory of the Competitive Firm under Price Uncertainty." *A.E.R.* 61 (March 1971): 65–73. (*b*)

Smith, Vernon L. "Tax Depreciation Policy and Investment Theory." *Internat. Econ. Rev.* 4 (January 1963): 80–91.

Sodersten, Jan. "Taxes and Investment Decisions." *Swedish J. Econ.* 74 (September 1972): 329–43

Theil, Henri. *Optimal Decision Rules for Government and Industry.* Amsterdam: North-Holland, 1964.

Bargaining, Yield Curves, and Wage Settlements: An Empirical Analysis

Yochanan Comay,* Arie Melnik,
and Abraham Subotnik

Technion—Israel Institute of Technology

We began by noting the lack of empirical work on bargaining, to partially redress this imbalance we drew a sample of strike-bound bargaining cases and translated each side's series of offers and counteroffers into monetary terms, thus creating management and union "yield" curves for each case. Some expectations as to the slope of these curves were gleaned from the literature and incorporated into three simple models—"linear," "quadratic," and "interactive " Separate ordinary least-squares regression of the offers and counteroffers for each side and each strike showed that the explanatory power of even such simple models was quite high, especially for management One result was that the union appears to be more responsive to the other side's offers, while management's behavior can be explained mainly by the passage of strike time

I. Introduction

The literature on bargaining models is by now profuse, integrated, and rigorous, though still lacking in empirical content [1] Some empirical progress has been made through experimental work by Siegel and Fouraker (1960) in their study of bargaining outcomes under various controlled conditions Empirical studies by Ashenfelter and Johnson (1969) on United States data, by Pencavel (1970) on Britain, and Vanderkamp (1970) on Canada have made casual use of the Hicksian bargaining model In comparison to the plethora of bargaining models, there does, however, appear to be an imbalance between model building and model testing in this field

---

* Dr. Comay died on active duty on November 24, 1973.

[1] Bargaining models have tended to develop by evolution, building one on another, and would make an interesting study in the history of ideas (see Zeuthen 1930, Nash 1950, Harsanyi 1956, Bishop 1964, Cross 1965, and Sarayder 1965)

In what follows we hope to partially redress this imbalance by examining the data from a sample of bargaining cases, where the bargaining was conducted during an ongoing strike, and by then testing a simple Hicks-like model Section I outlines Hick's model and its variant to be used here The survey of the literature that follows yields some expectations for such a model Section II deals with the data and the translation of offers and demands into monetary terms. Section III presents the regression results both for the different yield curves and for predicting the final settlement Results and conclusions are summarized in Section IV

II. The Model

Hick's bargaining model (1968), developed in the midthirties and reformulated in the sixties, has come under heavy attack over the years[2] His model consisted of two intersecting curves within a vertical wage-rate axis and a horizontal length-of-strike axis It was implied that the wage settlement would occur at the intersection of an "employer's concession curve" and a "union resistance curve" which were assumed to possess positive and negative slopes, respectively

For management, any point along the curve represents the highest wage they would be willing to pay rather than endure a strike of that length The opposite-sloped union-resistance curve is constructed similarly The final result is reproduced in figure 1 below

The z is the wage the employer would have paid if not subject to union pressure Hicks did not explicitly claim the point of intersection as the determinate bargaining solution, though the implication was clearly present

It can be shown that under a suitable reinterpretation, the point of intersection in figure 1 will in fact represent the point of static equilibrium, and that Hick's model does possess implications for the final settlement It does not, however, yield implications for the exact path of offers and counteroffers during the negotiation period beyond the result that the union's demand curve over time will follow a nonincreasing slope, while management's path will be nondecreasing[3]

There are several reasons which have been suggested elsewhere in the literature for expecting concessionary behavior, namely, a generally upward slope for management's offer curve and a downward slope for the union's curve over the length of strike Foremost among these is the mounting damage the strike entails in terms of loss of wages, loss of

---

[2] See, for example, Shackle (1964)

[3] As pointed out by a referee, the Hicksian model is phrased in terms of expected length of strike Bargaining during a strike can in fact be viewed as the process through which expected time influences real strike time.

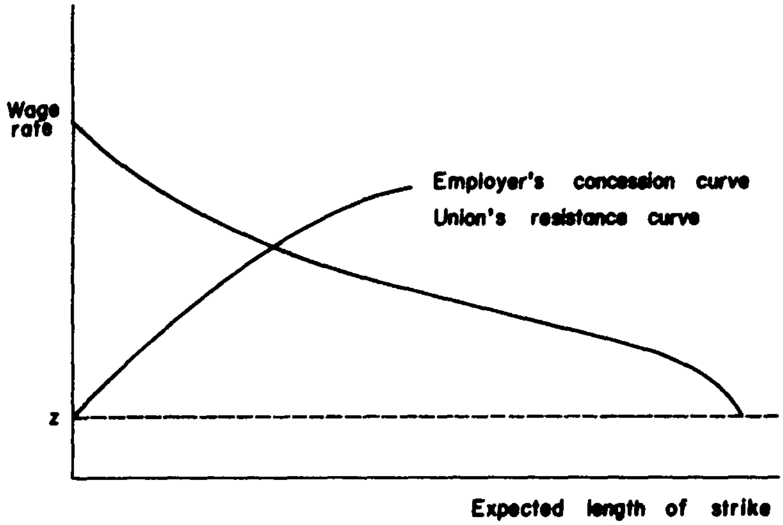


FIG 1

operating profits and market position, and damage to equipment through idleness [4] There is also a fixed recurring cost associated with the bargaining process itself [5]

Another reason which has been advanced for expecting concessionary behavior is the rise in the expected costs of continued disagreement as the strike progresses [6] Costs which had only been potential in the original prestrike reckoning now become a reality, also, as the strike progresses, the parties revise upward their expectations as to the length of strike associated with any persistent demand Most important, Chamberlain (1955) and Cartter (1959), as well as Schelling (1956), share the view that compromise behavior is quite rational, and is in fact a potent bargaining tool

The above discussion indicates that one may expect a positively sloped management curve and an opposite-sloped union curve with respect to strike time.[7] For simplicity we assume that the curves are linear.[8] For any given union-management bargaining case we thus have the following functions

---

[4] See Walton and McKersie (1965, p 31) Other losses enumerated by them are loss of management status with upper management or stockholders, loss of goodwill with labor, and loss of public image Equivalent measures on the union's side are loss of institutional security due to declines in membership, and loss of goodwill with management and with the public

[5] See Cross (1965, p 73)

[6] This point was made by Cartter, using a composite model incorporating elements of Chamberlain's and Pen's bargaining models It is worth noting parenthetically that Cartter dismisses estimation problems "since we do not anticipate an attempted empirical verification of this model of collective bargaining" (see Pen 1952; Chamberlain 1955, and Cartter 1959, pp 49–125)

$$U(S) = U(0) - \alpha t(S), \tag{1}$$

$$M(S) = M(0) + \beta t(S), \tag{2}$$

where $U(S)$ = a particular union's demand at the Sth bargaining session, with all elements of the demand translated into monetary terms; $U(0)$ = starting demand of the given union; $t(S)$ = time, in days, from start of strike to Sth session, α = rate of concession for the union, per unit of time, and $M(S)$, $M(0)$, and β = equivalent measures for management

Agreement, A, is reached at time $t(S^*)$ when the curves intersect, or $M(S^*) = U(S^*)$

$$\begin{aligned} U(0) - \alpha t(S^*) &= M(0) + \beta t(S^*) \\ t(S^*) &= \frac{U(0) - M(0)}{\alpha + \beta} \end{aligned} \tag{3}$$

Substitution of (3) in (2) yields the equation for the wage settlement

$$\begin{aligned} A &= \left(1 - \frac{\beta}{\alpha + \beta}\right) M(0) + \frac{\beta}{\alpha + \beta} U(0), \\ &= \frac{\alpha}{\alpha + \beta} M(0) + \frac{\beta}{\alpha + \beta} U(0) \end{aligned} \tag{4}$$

The A is thus a weighted average of the starting offers, with the weight dependent upon the rates of concession The higher α in relation to β, namely, the higher the relative rate of concession by the union, the closer the final settlement to management's position Similarity of behavior, namely, $\alpha = \beta$, will result in a settlement halfway between the starting positions

The assumption behind the linear model is that the costs per unit of time—or at least their subjective valuation—remain constant over the length of the strike This assumption could be unrealistic for several reasons For example, from the point of view of the union, the strike would entail a loss of wages; increasing marginal utility of income would make each passing strike day more costly. For management, the daily losses, in terms of income, customers, etc , should rise with the passage of strike time. The simplest method to incorporate this would be through quadratic yield curves, as in equation (5)–(6) below [9]

---

[7] The exact slopes will be determined by a host of factors On the union side, one could mention the size of the strike fund, the labor market situation, the perceived profit situation of the employer, public support for the strike, etc. Management's behavior will probably be influenced by its level of inventories before the strike, its financial rating, the structure of the market in which it operates, and the proportion of fixed costs out of total costs.

[8] Severe data limitations restricted the scope of the models which were empirically testable; the simplest forms were adopted in all cases, with a minimum number of parameters.

$$U(S) = U(0) - \gamma t^2(S), \tag{5}$$

$$M(S) = M(0) + \delta t^2(S) \tag{6}$$

These imply an increasing rate of concession by both parties

As noted by Cross (1965), however, the choice of which specific offer to make will be affected not only by the passage of time and its associated costs, but also by the counteroffer made by the other party In Cross's model, each side's offer affects the other's expectations as to his opponent's rate of concession Rapid (or slow) concessions by one side will cause the other to revise his expectations accordingly The level of his demands will be positively associated with his expectations as to the opponent's rate of concession, the opposite holds for the level of offers The conclusion Cross (1965, p 80) reaches is that "the greater is the concession rate of player II, the smaller will be the resulting concession rate of player I " In terms of our model, this implies a negative relation between α and β

As shown in table 1, this conclusion does not appear to be supported by the data, the relation between α and β is weakly positive [10] It appears that the yield behavior is different from that postulated by Cross, resulting in yield curves which are close to mirror images of each other In other words, higher concessions by one party are matched by higher concessions by the other, and vice versa The simplest manner of incorporating this element would be to make each side's counteroffer contingent on the other's prior offer, as is done in equations (7) and (8) below

$$U(S) = U(0) - \varepsilon t(S) M(S-1), \tag{7}$$

$$M(S) = M(0) + \zeta t(S)/U(S-1) \tag{8}$$

Equations (7) and (8) also imply an increasing rate of concession [11]

Solving as before, we arrive at the settlement function (9)

$$\begin{aligned} A &= \left[1 - \frac{\zeta/U(S^*-1)}{\varepsilon M(S^*-1) + \zeta/U(S^*-1)}\right] M(0) \\ &\quad + \left[\frac{\zeta/U(S^*-1)}{\varepsilon M(S^*-1) + \zeta/U(S^*-1)}\right] U(0), \\ &= \left[\frac{\varepsilon M(S^*-1)}{\varepsilon M(S^*-1) + \zeta/U(S^*-1)}\right] M(0) \\ &\quad + \left[\frac{\zeta/U(S^*-1)}{\varepsilon M(S^*-1) + \zeta/U(S^*-1)}\right] U(0) \end{aligned} \tag{9}$$

[9] More complex logarithmic forms were tried and found to be inferior in terms of explanatory power.

[10] The simple correlation between them is +.44.

[11] This can be seen by taking the second derivative of eqq. (7) and (8) with respect to time, t

III. Data

The New York State Bureau of Labor Statistics served as the source for our sample of bargaining cases which included only relatively large-scale strikes (1,500 workers or more), which occurred in the same period (1966–67), most of which took place in New York City Various issues of the *New York Times* were used to uncover information on the offer-counteroffer rounds during the strikes The cases finally selected for empirical analysis included only those for which five such rounds could be reconstructed, the cases are listed in the Appendix

All offers and counteroffers were converted into monetary terms and expressed as percentages of the prestrike mean wage[12] In converting nonwage items into monetary terms, certain guidelines were followed Reductions in work time were considered equivalent to wage increases For example, a decrease of the working week from 40 to 36 hours, with the weekly wage remaining unchanged, was evaluated as a 10 percent wage increase Pensions and special retirement clauses, which involve futurity, were discounted to the strike period using a discount rate of 4 percent and 20 years of remaining work time[13] Where negotiations were conducted over future wage increases to be granted during the period of the contract, these were also discounted to the year of bargaining All offers and counteroffers were summarized, therefore, as percentages of current wages

IV. Results

To test the adequacy of the above formulations, we restricted our sample to strikes which encompassed five or more offer-counteroffer rounds between the start of the strike and its settlement. The final sample included eight bargaining cases The results of ordinary least-squares regression of the union and management offers for each strike according to the linear, quadratic, and interactive models of equations (1), (2), (5), (6), (7), and (8) are summarized in table 1

All three models perform better in explaining management's offer curve than in explaining the union's With the exception of the television artists' strike all coefficients of determination for management are .80 or higher For the union, on the other hand, in four out of the eight strike cases the coefficients of determination are below 61

---

[12] Mean wage was calculated as the mean of beginning and senior workers' salaries, since no information was available on the seniority distribution of union members.

[13] Estimates were relatively insensitive to differing assumptions concerning the remaining length of work time Estimates were more sensitive to variations in discount rates, however, due to the relative insignificance of the items concerned, the final percentage estimate was not altered significantly by changes in the discount rates.

TABLE 1
REGRESSION RESULTS OFFER AND COUNTEROFFER BY STRIKING PARTY

	MANAGEMENT						UNION[c]					
	Linear Model		Quadratic Model		Interactive Model		Linear Model		Quadratic Model		Interactive Model	
STRIKE[a]	Coefficient[b]	R^2	Coefficient	R^2	Coefficient	R^2	Coefficient	R^2	Coefficient	R^2	Coefficient	R^2
1 Ford	052**	99	001**	99	0 332**	99	007	63	0001*	66	0014*	75
2 Marine	624*	91	112**	95	20 584**	91	356	36	067	40	0152	38
3. Five airlines	149**	95	004**	89	2 361**	95	035**	93	0008**	84	0025**	91
4 Board of Education	235**	95	014**	84	6 665**	94	165*	33	012*	44	0095*	35
5. Pan American	484*	80	.050**	89	6 776*	80	040	30	004	35	.0067	.61
6 Television	099*	40	011**	59	1 952**	45	136**	44	015**	61	0128**	.54
7 Moving	778**	96	059**	91	14 215**	96	345**	75	029**	84	0449**	.94
8 Builders	249**	94	008**	94	7 546**	98	562**	97	018**	95	0308**	98

[a] A fuller identification of the striking parties is available in the Appendix

[b] Standard deviations have been omitted to economize on space

[c] The negative sign in front of all coefficients has been omitted to economize on space

* p > .05

** p > 01

The quadratic model appears to be superior to the linear in explaining the variance of the union offers Thus, in four out of the eight cases, the percentage of variance explained by this model is higher by 5 percent or more, while in three cases the two models are similar The union tends to hold out for long periods and then yield rapidly toward the end of the strike, thus creating concavity in its yield curve What is most interesting, however, is that the introduction of interactive elements in the union's function improves the fit in four of the bargaining cases, in the rest the interactive model is about the same in explanatory power as the others

The same cannot be said for management, where the interactive model does not improve the fit significantly In fact, all three models perform about the same, except in the television artists' case, where the quadratic is clearly superior Thus, there is some evidence that in management's case primarily the passage of strike time explains their offer behavior, while the union is more responsive to the actual offers made by management

The accuracy of the models in actually predicting the final settlement can be assessed by solving for these values according to the three models, utilizing the results from table 1, It is unclear a priori which will perform better, since this test incorporates both blades of the scissors, union and management The results are summarized in table 2.

The superiority of the interactive model in predicting the final settlement is evident from the table In half of the cases it improves the prediction substantially, while in the other half it does as well as the linear Its relative success is not surprising in light of its greater utilization of available information Thus, for each round, the model uses information on the other party's behavior in the previous round to predict the current offer What is somewhat surprising is the closeness of the predicted and actual outcome, even for such a skeletal model In absolute terms the greatest deviation is just under 5 percentage points, for the Board of Education strike In this case the union set its demand on the first day of the strike and did not budge through 11 bargaining rounds—while the Board of Education yielded continuously The other cases range between zero and 1 8 percentage points As a percentage of the actual settlement this means a deviation of 26 percent for the Board of Education, and 0–7 percent for the other cases

V. Summary and Conclusions

We began by noting the lack of empirical work on bargaining, especially in light of the vast array of extant theoretical models. To partially redress this imbalance, we drew a sample of strike-bound bargaining cases and translated each side's series of offers and counteroffers into monetary terms, thus creating management and union "yield" curves for each case

TABLE 2
ACTUAL AND PREDICTED SETTLEMENTS, BY STRIKING PARTY

STRIKE AND MODEL[a]	SETTLEMENT (%)		
	Actual	Predicted	Difference
Ford			
L	5 8	6 1	0 3
Q		3 6	−2 2
I		6 1	0 3
Marine			
L	27 0	28 9	1 0
Q		26 1	−0 9
I		28 8	1 8
Five airlines			
L	15 4	15 5	0 1
Q		10 1	−5 3
I		15 3	−0 1
Board of Education			
L	18 8	23 6	4 8
Q		15 6	−3 2
I		23 7	4 9
Pan American			
L	12 5	13 3	0 8
Q		5 7	−6 8
I		12 7	0 2
Television			
L	14 2	14 6	0 4
Q		15 9	1 7
I		14 2	0 0
Moving			
L	12 2	14 6	2 4
Q		7 0	−5 2
I		12 9	0 7
Builders			
L	24 0	22 5	−1 5
Q		35 5	11 5
I		23 3	−0 7

[a] L = "linear," Q = "quadratic," and I = "interactive "

Some expectations as to the slope of these curves was gleaned from the literature and incorporated into three sample models—"linear," "quadratic," and "interactive "

Despite the limited nature of the sample, there are several conclusions which emerged even from an analysis of the raw data Thus, the yield behavior of the union appears to be different in nature from that of management, and consists of a lengthy period of little or no concession followed by rapid yielding in the last few days of the strike. Management yields gradually along the period of the strike

For both parties, however, the rate of concession increases during the strike period, resulting in concession curves which are concave to the origin This could stem from the rise in the strike cost per day for both parties as the strike progresses. One possible explanation for the union could relate to the decline in the strike fund and the diminishing opportunities for union members to purchase on credit. For management, the losses increase not only through loss of sales but also through idle and deteriorating equipment, loss of potential clients, etc. The other explanation put forward in this paper is that the concavity also stems from interactions between the bargaining parties As was noted, the resultant shape of the yield curves does not appear to be consistent with Cross's bargaining model.

Separate ordinary least-squares regression of the offers and counteroffers for each side and each strike showed that the explanatory power of even such simple models was quite high, especially for management One interesting result was that the introduction of interaction into the model improved the specification for the union, while not significantly affecting the results for management The union thus appears to be more responsive to the other side's offers, while management's behavior can be explained mainly by the passage of strike time We then compared the models from the point of view of predicting the final settlement, here the superiority of the "interactive" model emerged most clearly

Appendix

List of Bargaining Cases[14]

1 *Ford Motor Co* vs *United Auto Workers* (September 7, 1967)

2 *Marine Touring and Transportation Association* vs. *Maritime United Marine Division* (*Local 333, AFL-CIO*) (January 30, 1967)

3 Five Airlines (*United, TWA, Pan-Am, Eastern, National*) vs *Machinists Union, AFL-CIO* (July 8, 1967)

4 *New York City Board of Education* vs. *Teachers Union* (*Local 2*) (September 11, 1967)

5. *Pan American World Airways* vs *Air Line Pilots Association* (March 31, 1965)

6 *ABC, CBS, and NBC* vs *Federation of Television and Radio Artists* (March 29 1967).

7 *Moving and storage companies in New York City and vicinity* vs. *International Brotherhood of Teamsters* (March 1, 1965).

8 *Southern Tier Builders Association*, Binghamton, New York vs. *Bricklayers, Masons, and Plasterers Int* (May 1, 1965)

References

Ashenfelter, O., and Johnson, G R. "Bargaining Theory, Trade Unions, and Industrial Strike Activity." *A.E.R* 59 (March 1969). 35–49.

Bishop, R. "A Zeuthen-Hicks Model of the Bargaining Process." *Econometrica* 32 (July 1964)· 411–17.

[14] Date in parenthesis refers to the first day of each strike

Cartter, A. M. *Theory of Wages and Employment.* Homewood, Ill.: Irwin, 1959.

Chamberlain, N. W. *A General Theory of Economic Process.* New York. Harper & Row, 1955.

Cross, J. G. "A Theory of the Bargaining Process." *A E.R.* 55 (March 1965): 67–94.

Harsanyi, J. C. "Approaches to the Bargaining Problem Before and After the Theory of Games." *Econometrica* 24 (April 1956) · 144–57.

Hicks, J R *The Theory of Wages.* 2d ed London St Martin Press, 1968.

Nash, J. F. "The Bargaining Problem " *Econometrica* 18 (April 1950): 155–62.

Pen, J. "A General Theory of Bargaining " *A E.R.* 42 (March 1952) · 24–42.

Pencavel, J. H "An Investigation into Industrial Strike Activity in Britain." *Economica* (August 1970), pp. 239–56

Sarayder, E "Zeuthen's Theory of Bargaining· A Note." *Econometrica* 33 (December 1965). 803–13.

Schelling, T C "An Essay on Bargaining " *A.E.R* 46 (June 1956) 281–306

Shackle, G L. S "The Nature of the Bargaining Process " In *The Theory of Wage Determination,* edited by J T. Dunlop London Macmillan, 1964.

Siegel, S , and Fouraker, E *Bargaining and Group Decision Making* New York. McGraw-Hill, 1960

Vanderkamp, J "Economic Activity and Strikes in Canada " *Indus Relations* 9 (February 1970) 215–30.

Walton, R E , and McKersie, R B *A Behavioral Theory of Labor Negotiations.* New York McGraw-Hill, 1965

Zeuthen, F. *Problems of Monopoly and Economic Welfare* London Routledge, 1930

Interpretations of Departures from the Pareto Curve Firm-Size Distributions

---

Yuji Ijiri and Herbert A Simon

Carnegie-Mellon University

Empirical firm-size data show concavity when size is plotted against rank (largest firm = rank 1) on a log-log scale, a departure from the Pareto distribution which shows a straight line. Two reasons for this departure are proposed One is an exponential decay in the effect of a unit increase in size upon the firm's future growth It is shown analytically that the greater the decay rate, the greater the concavity. The other is an effect of mergers and acquisitions Based on Federal Trade Commission data, the actual firm-size distribution in 1969 is compared with a hypothetical one where all mergers and acquisitions in 1948–69 were "undone" to demonstrate their effect

I. Concavity of the Firm-Size Distributions

It has been observed repeatedly that the size, s, of a firm and its rank, r, in an industry or in an economy satisfy approximately the Pareto law (see, for example, Steindl, 1965)

$$sr^{\beta} = A, \tag{1}$$

where β and A are constants The size may be measured by annual sales, otal assets, or the number of employees The rank is based on the selected ize measure, the largest firm being assigned rank 1.

In a previous paper (Ijiri and Simon 1971), we used the constant β as in indicator of the degree of concentration in the population because it ndicates the frequencies of large firms relative to smaller firms. For xample, if the rank of a firm X is twice the rank of a firm Y, the size of X is $2^{-\beta}$ times the size of Y, since $s_X = Ar_X^{-\beta} = Ar_Y^{-\beta}2^{-\beta} = 2^{-\beta}s_Y$ The larger the β, the greater the relative size of a large firm (small rank)

The authors are professor of industrial administration and R. K. Mellon Professor of omputer Science and Psychology, respectively, at Carnegie-Mellon University This roject has been supported in part by a grant from the National Science Foundation.

compared with a smaller firm (large rank). Clearly, A is the size of the largest firm

When $\log s$ is plotted against $\log r$, the relation is a straight line·

$$\log s = \log A - \beta \log r \tag{2}$$

Thus, β is the slope of the line on a double-log scale.

The empirical data on firm sizes fit this distribution reasonably well as a first approximation (Steindl 1965) However, we often encounter firm-size distributions for the economy that have some concave (downward) curvature For example, the *Fortune* data for 1969 (*Fortune* 1970*a*, 1970*b*) show a considerable curvature for the size distribution based on amount of assets (total assets after accumulated depreciation) for the 831 largest industrial firms This distribution is shown in figure 1 [1] The straight line through the empirical data is the Pareto curve fitted to the tenth largest firm and the eight hundred thirty-first firm [2] (The first nine firms are ignored in fitting the curve in order to take advantage of the statistical averaging incorporated in the cumulative distribution function except at its extreme tail)

As shown in figure 1, the actual size of a firm around rank 100 is almost twice as large as the "theoretical" size predicted from the straight line That is, actual assets of the one hundredth firm are $1.01 billion while "theoretical" assets are $0 50 billion Of course, many lines can be drawn on the figure as "theoretical" distributions The above illustration is intended only to show the significance of the curvature of the actual firm-size distributions [3]

The ratio of the actual size to the theoretical size of a firm with a given rank may be called a "size variance " Then, the downward concavity of the distribution shows an upward size variance for middle-rank firms relative to small- or large-rank firms, where the size variance is measured from the Pareto curve

Faced with this rather significant departure from the theoretical distribution, one might abandon the Pareto curve and look for other distributions which, when the cumulative frequencies are plotted on a double-log scale, show curvature similar to the empirical distributions [4]

---

[1] Empirical size-rank distributions based on sales size and on assets size are almost indistinguishable except for the difference in the intercept as can be seen by plotting data from *Fortune* (1970*a*, 1970*b*).

[2] It has been shown (Simon 1955) that the Pareto curve can be derived from Gibrat's law, which states that the percentage growth rate of a firm is distributed independently of its size

[3] The selection of the tenth and the eight hundred thirty-first firms is arbitrary However, we use the "theoretical" distribution only for purposes of illustration. Our conclusions, especially with regard to the effect of mergers and acquisitions on the size distribution, would hold equally well for other methods of fitting a Pareto curve to the data. We have therefore used the simplest possible method of fitting—drawing the curve through two points in the cumulative distribution.

[4] The lognormal and negative binomial are examples of such distributions.

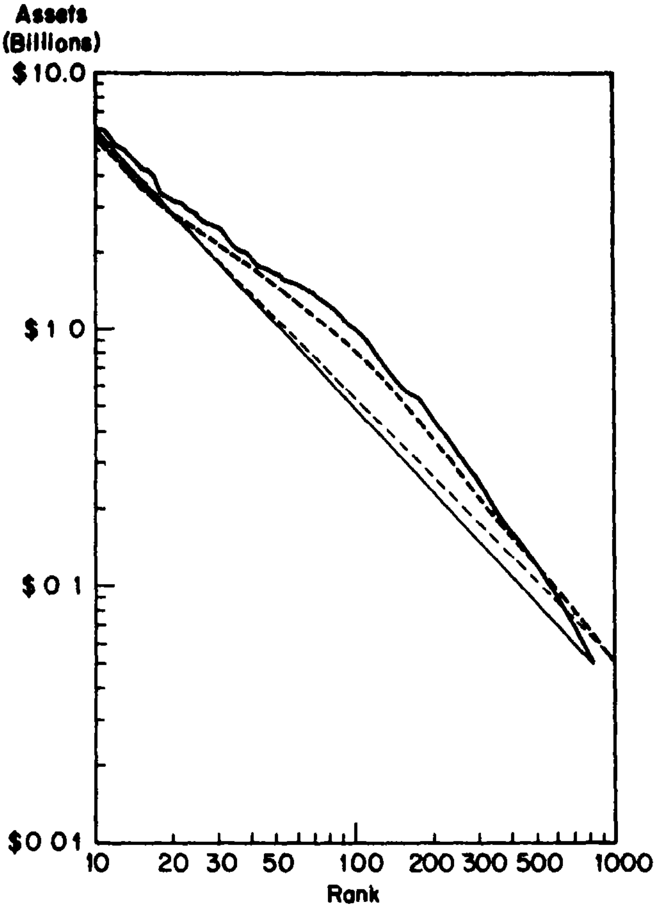


Fig 1 —Solid curve actual size-rank relationship as of December 31, 1969 Broken urve estimated size-rank relationship as of December 31, 1969, when the effects of nergers and acquisitions in 1948–69 were eliminated. Solid line theoretical size-rank elationship (the Pareto curve) for actual data (connecting the sizes of the tenth and ight hundred thirty-first firms) Broken line same as solid line for estimated "mergerless" lata

Iowever, the Pareto curve is derivable from simple and economically lausible assumptions as shown in Simon (1955), namely, size indepenlence of percentage growth rate (Gibrat's law) and constancy of the ntry rate Both assumptions have significant economic implications Ience, it seems to be profitable to attempt to retain the Pareto curve as a irst approximation and to explain the deviations from this distribution y considering additional economic factors that violate the assumptions f the derivation, and hence could produce perturbations.

The purpose of this paper is, therefore, to explore the ways of explaining the size variance by economic processes that can be observed empirically. We consider two alternative explanations. The first extends an earlier paper (Ijiri and Simon 1964), which analyzed by simulation the effects of autocorrelation of growth, giving a firm with a history of recent growth a better chance for further growth than a firm of the same size whose growth had taken place in the distant past.

Under Gibrat's law, a unit increase in the size of a firm contributes to its future growth, since the base to which the percentage growth rate is applied is thereby increased by one unit. But the effect of a unit increase most likely diminishes as time goes on. In the simulation model in Ijiri and Simon (1964), this pattern of diminishing effects is represented by an exponential decay. The simulation model showed that the greater the decay (or discount) rate, the greater the concavity of the distribution. This decay rate has a significant economic meaning: it measures the effect of growth momentum and of the stability of the economic environment. For example, when consumer brand loyalty is high and market share is stable, recent growth has a long carry-over effect. When consumer brand loyalty is low and market shares are unstable, recent growth has only a short-run effect upon the firm's future growth.

We are now able to derive analytically this effect of the decay rate, which we estimated by simulation in our earlier paper. In the next section, we summarize our analytical characterization of this model, leaving mathematical derivations to the Appendix.

In Section III, we consider a second economic process that may contribute to the concavity of the distribution, namely, the effect of mergers and acquisitions. In an earlier paper (Ijiri and Simon 1971), we analyzed the effect of mergers and acquisitions on the concentration measure β. When we wrote that paper, we had available only a very limited amount of data on mergers and acquisitions. We now have analyzed a more extensive set of data over a wider period, based on the Federal Trade Commission's report covering mergers and acquisitions of manufacturing and mining companies between 1948 and 1969. Contrary to our previous findings, the results of this analysis show that mergers and acquisitions have contributed to a noticeable extent to increasing the concentration measure and increasing the concavity of the firm-size distribution. We present our results of the merger analysis in Section III.

II. Autocorrelated Growth

The curvature of a firm-size distribution may be quantified in many different ways when we fit a curve to empirical data. It is possible to fit

$$\log s = \log A - \beta \log r + c(\log r)^2, \qquad (3)$$

where the coefficient c depends on the curvature The distribution is convex downward if c is positive and concave downward if c is negative

Alternatively, we might select a set of pairs of ranks i and j and measure a representative value (mean, median, etc) of the ratios for each pair

$$d = s_k/(s_i^{\upsilon} s_j^{1-\upsilon}), \tag{4}$$

where s_k is the size of the rank k firm $(i < k < j)$ and υ satisfies

$$k = i^{\upsilon} j^{1-\upsilon} \tag{5}$$

The denominator in (4) is the theoretical value for the size of the kth firm under the Pareto law, and the numerator is its actual value

A stronger theoretical basis for quantifying and interpreting the curvature can be obtained by using a parameter that we introduced in an earlier paper (Ijiri and Simon 1964) In that paper, we demonstrated by a simulation that the following process leads to a distribution which approximates the Pareto curve but with a curvature

1 At the start of the process, namely, at epoch 0 (a specific point in time), we have several firms in the population (an industry or an economy), each firm having a unit (sales) size[5] and weight 1 The weight is an indicator of the firm's growth potential relative to other firms Placing several firms in the population is an initialization step for the model It has been shown that after large numbers of epochs the resulting distribution is insensitive to the initialization (see Ijiri and Simon 1964)

2 At each epoch t, $t = 1, 2, 3,$, one unit of sales is given to a randomly selected firm

3 There is a constant probability α that the sales unit at epoch t goes to a new firm If this happens, a new firm with size 1 is added to the population This firm is treated as an old firm in all succeeding epochs

4 If the sales unit at epoch t does not go to a new firm, it goes to an old firm The probability of a given old firm receiving the sales unit is proportional to its weight—that is, it is equal to the firm's weight divided by the sum of weights of all firms in the population The size of the selected firm is increased by one unit

5 After a new firm is created or an old firm is selected among firms in the population, the weight of each old firm in the population is discounted by multiplying its current weight by γ Then, if a new firm is created, its weight is set equal to 1, if an old firm is selected, its weight (after the above discounting) is increased by 1

In our earlier paper (Ijiri and Simon 1964), we showed by simulation hat the resulting firm-size distribution exhibits a downward concavity if < 1, is a straight line if $\gamma = 1$, and has a downward convexity if $\gamma > 1$ The curvature is greater as $|\gamma - 1|$ is greater. (For the case of $\gamma = 1$, namely, without discounting, see Simon 1955) Therefore, γ may be used to quantify the curvature observed in an empirical distribution by fitting a theoretical distribution generated under a given γ

---

[5] This can be stated in terms of a unit of assets, but we shall use sales units for the ake of ease in explanation.

The parameter γ has a simple economic interpretation It indicates the relative significance of recent sales as compared with earlier sales in generating future sales Both recent sales and prior sales contribute in generating future sales, but the latter contribute less than the former if $\gamma < 1$ and more than the former if $\gamma > 1$ Thus, if $\gamma < 1$, a firm experiencing growth recently has a better chance of growing than a firm of the same size that had its growth in earlier periods

We now show these points analytically for the case where $0 < \gamma < 1$ (distribution concave downward)

Let W_t be the sum of weights of all firms in the population after step 5 for epoch t has been completed Then,

$$W_{t+1} = \gamma W_t + 1, \tag{6}$$

as described in step 5 Then, for a sufficiently large t,

$$W_t = 1 + \gamma + \gamma^2 + \qquad + \gamma^t \cong \frac{1}{1 - \gamma} \; (0 < \gamma < 1) \tag{6a}$$[6]

Let us then consider a firm that was created at epoch t The size of this firm at epoch t is 1 and so is its weight Then, the probability that this firm will experience a sale at epoch $(t + 1)$ is the probability that the sale goes to an old firm $(1 - \alpha)$ times the probability that this particular firm experiences the sale, $[1 - 1/(1 - \gamma) = 1 - \gamma]$, or a total probability, $(1 - \alpha)(1 - \gamma)$, which we denote by δ Thus, the probability that this firm does not get the sale at epoch $(t + 1)$ is $(1 - \delta)$

Now assume that this firm did not get the sale at epoch $(t + 1)$ The weight is now discounted from 1 to γ for epoch $(t + 2)$ Thus, the probability of getting the sale at epoch $(t + 2)$ is $(1 - \alpha)\,(1 - \gamma)\,\gamma = \delta\gamma$, while the probability of not getting the sale at epoch $(t + 2)$ is $(1 - \delta\gamma)$ Similarly, the probability of not getting the sale at epoch $(t + 3)$ is $(1 - \delta\gamma^2)$

Therefore, the probability that this firm will not get any sale after epoch t (that is, the probability u_1 that the size of this firm remains at 1 forever) is

$$u_1 = (1 - \delta)(1 - \delta\gamma)(1 - \delta\gamma^2)\ldots \tag{7}$$

---

[6] More precisely, $W_t = 1 + \gamma + \gamma^2 + \cdots + k\gamma^t$, where k is the number of firms created at initialization with size 1 and weight 1. Clearly, k has only a small effect when t is large.

For γ close to but less than 1, $\delta = (1 - \alpha)(1 - \gamma)$ is close to 0, and so is $\delta\gamma^t$ for $t = 1, 2, \ldots$[7] Then, we may use the approximation

$$\ln(1 + x) \cong x, \tag{8}$$

obtaining

$$\ln u_1 \cong -\delta(1 + \gamma + \gamma^2 + \cdots) = -(1 - \alpha)(1 - \gamma)/(1 - \gamma) \tag{9}$$

or

$$u_1 \cong e^{-(1-\alpha)} \tag{10}$$

In general, the probability of a firm with a present weight w not getting any sales in the future is easily shown to be

$$u_w \cong e^{-(1-\alpha)w} \tag{11}$$

by applying the above argument Equation (10) is a special case for $w = 1$

The implication of the expression in (11) to the firm-size distribution is this First, u_1 is the expected asymptotic proportion of the unit-sized firms among all firms In (10), u_1 depends only on α and is independent of γ

Next, consider a firm of size s and weight w Here,

$$w = \sum_{k=0}^{\infty} \lambda_k \gamma^k, \tag{12}$$

where $\lambda_k = 1$ if the firm received the sale at epoch $(t - k)$ and $\lambda_k = 0$ otherwise The sequence $\lambda_0, \lambda_1,$, should have exactly s 1's Therefore, for any given history of a firm (given $\{\lambda_k\}$),

$$\frac{dw}{d\gamma} > 0. \tag{13}$$

But, the probability of this firm not receiving any more sales is given by (11), namely, $u = (1 - \delta w)(1 - \delta\gamma w)(1 - \delta\gamma^2 w) \quad \cong e^{-(1-\alpha)w}$ Hence,

$$\frac{du}{dw} = \frac{du}{d\ln u}\frac{d\ln u}{dw} = u \sum_{t=0}^{\infty} - \delta\gamma^t/(1 - \delta\gamma^t w) < 0 \tag{14}$$

Thus, from (13) and (14)

$$\frac{du}{d\gamma} < 0 \tag{15}$$

This means that the probability of having further sales $(1 - u)$ for old firms becomes smaller as γ becomes smaller (namely, as discounting becomes heavier). Growth beyond size 1 becomes more and more difficult as weights are discounted more heavily (or as γ becomes smaller).

---

[7] If $\gamma(<1)$ is not close to 1, $\delta\gamma^t$ becomes close to zero except for the first few terms. In the simulation in Ijiri and Simon (1964), it was noted that even a mild departure of γ from 1, such as 0.95, is sufficient to create a considerable curvature in the distribution.

Furthermore, it can easily be seen from (12) that on the average as the size s of a firm increases, the effect of reducing γ on w becomes more severe since there are more terms in the expression for w that are reduced as γ becomes smaller Thus, on the average, reducing γ has a greater effect on the probability of further growth as size s becomes larger Therefore, not only is the probability of further growth for old firms decreased as γ becomes smaller, but also the decrease in probability is greater the larger the firm

From this analysis, we can conclude that as γ becomes smaller the concavity of the firm-size distribution on a double-log scale is increased

In the Appendix, we present a further analysis of the effect of γ on the concavity from the viewpoint of the mean and the variance of the distribution of firm sizes

The empirical support for this autocorrelation between recent growth and future growth (in our analysis this means $\gamma < 1$) has been given in Ijiri and Simon (1967) There, an excess growth ratio of a firm in 1954–58 over and above the average growth ratio of the economy is definitely carried over to the subsequent period in 1958–62 but only in proportion to the cube root of the original excess growth ratio

Empirically, this decay effect appears to be quite plausible When two firms of equal size are compared, the one that grew more recently seems to have a better chance for growth than the other whose growth occured in the remote past Growth momentum is not likely to change suddenly

Thus, the existence of concavity in the firm-size distribution does not seem to be a good reason for abandoning the Pareto curve but a good reason for modifying the Pareto curve by introducing the autocorrelation in the growth process

III. Mergers and Acquisitions

Let us now turn to the second economic factor that may have contributed to creating concavity in the firm-size distribution, namely, the effects of mergers and acquisitions

In our previous paper (Ijiri and Simon 1971), we analyzed the steady-state distribution of firm sizes in an economy in the following two steps A first process selects firms that disappear as a result of mergers and acquisitions Then, the assets of all disappearing firms are pooled and allocated to surviving firms. This separation into two steps allowed us to analyze data that were not detailed enough to make a correlation analysis of the relation between the sizes of acquiring firms and the sizes of acquired firms

The theoretical model to which the data were compared in Ijiri and Simon (1971) was based on the following two assumptions (1) the probability of a firm disappearing is independent of its size, and (2) the pooled assets of disappearing firms are allocated among surviving firms in such a way that the percentage increase due to allocation in the pooled size of the surviving firms in a given size range is independent of the size range The empirical data we examined in Ijiri and Simon (1971) were not extensive enough to detect any significant departure from the above assumptions These data referred to mergers and acquisitions that occurred in 1956–57, when mergers and acquisitions were not as frequent as in the 1960s

We obtained and examined more detailed data covering the period 1948–69 [8] After having analyzed this new set of data to retest the model, we found significant departures from the above assumptions, sufficient to produce some curvature of the firm-size distribution

Tables 1 and 2 are based on the Federal Trade Commission's publication *Large Mergers in Manufacturing and Mining, 1948–1969* (1970) and the *Fortune* data (1970*a*, 1970*b*) on assets of the largest industrial corporations as of December 31, 1969 In table 1, we assemble the basic data needed for estimating the "merger-free" distribution—the distribution which would be obtained from the actual 1969 data had no mergers and acquisitions occurred in the 1948–69 period We can then see the effect of mergers and acquisitions on the firm-size distribution by comparing the (estimated) merger-free and the (actual) postmerger distributions Table 1 also tests the assumption (1 above) that the probability of disappearance of firms by mergers and acquisitions is independent of the size

In table 1, column 1 shows the frequencies of firms in the indicated asset size ranges as of December 31, 1969, based on the *Fortune* data Columns 2, 3, and 4 show the frequencies of firms involved in mergers and acquisitions in 1948–69 as reported in the Federal Trade Commission report (1970) They are classified in three ways, based on postmerger size (col 2), the premerger size of the acquiring firm (col 3), and the premerger size of the acquired firm (col 4)

Now suppose that all mergers in 1948–69 had occurred at once just

---

[8] There are a number of differences between the data in this present study and those in the previous one (Ijiri and Simon 1971) that complicate our task of interpreting the findings The present study covers a longer time period and more firms, enabling us, for example, to classify firms with 10 size categories instead of the four categories of the earlier study. Size is measured in the present study by assets and in the previous study by sales, but we do not believe that this difference changes matters very much. It is entirely possible that the Pareto curve provides a better approximation for the earlier period both because mergers were less frequent and because those that occurred were distributed in closer conformity to the assumptions of the model than the more frequent mergers of the 1960s were.

TABLE 1
FREQUENCIES OF FIRMS CLASSIFIED BY ASSET SIZES

SIZE CLASS AND ASSET SIZE RANGE ($ BILLION)	NUMBER OF LARGE INDUSTRIAL FIRMS IN *Fortune* 1,000 IN 1969 (1)	NUMBER OF FIRMS INVOLVED IN MERGERS, 1948–69, CLASSIFIED BY			ESTIMATED NO OF FIRMS IF EFFECTS OF MERGERS ARE ELIMINATED (1) − (2) + (3) + (4) (5)	RATE OF DISAPPEARANCE (4)/(5) = (6)
		Postmerger Size of Combined Firms (2)	Premerger Size of Acquiring Firms (3)	Premerger Size of Acquired Firms (4)		
1) Over 12.8	2				2	0.00
2) 6.4–12.8	7	4	4		7	0.00
3) 3.2–6.4	11	14	11		8	0.00
4) 1.6–3.2	33	36	31	1	29	0.03
5) 0.8–1.6	70	94	83	4	63	0.06
6) 0.4–0.8	95	147	127	9	84	0.11
7) 0.2–0.4	137	184	172	20	145	0.14
8) 0.1–0.2	214	206	171	83	262	0.32
9) 0.05–0.1	262	159	160	139	402	0.35
	831	844	759	256	1,002	
10) Below 0.05		146	231	734		
		990	990	990		

TABLE 2
AGGREGATE ASSETS OF FIRMS CLASSIFIED BY ASSET SIZES

SIZE CLASS AND ASSET SIZE RANGE	*Fortune* 1,000 IN 1969 (1)	FIRMS INVOLVED IN MERGERS, 1948–69			ESTIMATED WITHOUT MERGERS (1) − (2) + (3) + (4) (5)	SURVIVING FIRMS (5) − (4) (6)	ACQUIRED FIRMS' ASSETS CLASSIFIED BY SIZE OF ACQUIRING FIRMS (7)	RATE OF GROWTH (7)/(6) (8)
		Post-merger Firms (2)	Ac-quiring Firms (3)	Ac-quired Firms (4)				
1) Over 12.8	32.36				32.36	32.36		0.00
2) 6.4–12.8	54.61	39.09	38.89		54.41	54.41	0.20	0.00
3) 3.2–6.4	49.86	56.87	45.20		38.19	38.19	0.91	0.02
4) 1.6–3.2	72.56	78.03	69.32	1.85	65.70	63.85	4.95	0.08
5) 0.8–1.6	82.07	106.66	93.65	4.23	73.29	69.06	8.52	0.12
6) 0.4–0.8	53.23	83.78	71.97	5.24	46.66	41.42	10.28	0.25
7) 0.2–0.4	38.71	54.67	50.74	6.11	40.89	34.78	11.43	0.33
8) 0.1–0.2	29.69	30.76	24.95	11.39	35.27	23.88	7.77	0.33
9) 0.05–0.1	18.54	12.21	11.80	9.53	27.66	18.13	5.12	0.28
	431.63	462.07	406.52	38.35	444.43	376.08	49.18	
10) Below 0.05		5.51	6.18	16.53			5.70	
		467.58	412.70	54.88			54.88	

before the close of the year 1969. This is not a totally unrealistic assumption, since by far the greatest portion of the mergers are concentrated in the last 2–3 years of the 20-year period. We can then calculate the size distribution of firms that would result if all such mergers and acquisitions were "undone" at a single moment in time, as if the merger movie were suddenly and swiftly run backward.

Take, for example, the size class 5 of firms with assets of $0.8–$1.6 billion. As of December 31, 1969, the *Fortune* data show 70 firms in this size class. We want to estimate what this number would be if mergers and acquisitions were undone all at once. To do so, we first note that in this size class there were 83 instances where firms acquired other firms and four where firms were acquired by other firms. On the other hand, there were 94 merger and acquisition instances where firms, after merger, belonged to this size class. Hence, there was a net gain of seven (94 − 83 − 4 = 7) in the number of firms in this class by reason of mergers and acquisitions. Since the actual number of firms in this class in 1969 is 70, we estimate that 70 − 7 = 63 would have been in existence in 1969 had there been no mergers, as stated in column 5. By taking the cumulative frequencies based on the data in column 5 and plotting the result in figure 1 as a broken curve, we obtained an estimated size-rank relationship when all mergers and acquisitions are "undone." We shall come back to this point again shortly.

Finally, the last column, column 6, shows the percentage rate of disappearance of firms in each size class by mergers and acquisitions over the 20-year period. From this we can conclude that smaller firms have a higher chance of being absorbed by mergers and acquisitions, contrary to the assumption (1) of equal probability of disappearance.

To test the assumption (2), on the size independence of growth rate by mergers, we prepared table 2. Column 1 was taken from the *Fortune* data, adding the assets of all firms belonging to a given size class. Columns 2, 3, and 4 were prepared in the same way as in table 1 except that they refer to aggregate assets instead of frequencies.

Consider, again, the size class 5. The table shows that the aggregate assets of the postmerger firms belonging to this class are $106.66 billion while the aggregate assets of the acquiring and acquired companies are $93.65 and $4.23 billion, respectively. Thus, as a result of mergers and acquisitions, this size class gained $8.78 billion in aggregate assets (= 106.66 − 93.65 − 4.23). Since the actual aggregate assets based on the *Fortune* data are $82.07 billion, we estimate the "mergerless" aggregate size of this class to be $73.29 billion, as shown in column 5. Other figures in this column were prepared in the same way.

From data in column 5, we eliminated the aggregate assets of acquired firms belonging to the class (col. 4). The result is shown in column 5. We would like to see how these aggregate assets of acquired firms are redistributed among the surviving firms. Column 7 shows the distribution of the aggregate assets of acquired firms based on the premerger size of the *acquiring* firms by which the acquired firms were absorbed Thus, for example, the size class 5 increased its aggregate assets by 12 percent as a result of reallocation of assets of acquired firms From the data in the last column, it seems clear that the assumption (2) of size-independent growth rate by reallocation is not satisfied [9]

What then is the effect of mergers and acquisitions on the firm-size distribution? To show this effect graphically, we prepared from column 5 of table 1 the broken curve in figure 1 The broken curve indicates an estimated firm-size distribution when the effects of mergers and acquisition are eliminated This was superimposed on the actual postmerger size distribution given by the solid curve in figure 1

The broken line was obtained by connecting the sizes of the tenth and eight hundred thirty-first firms in the "mergerless" distribution in order to make it comparable to the solid line indicating the theoretical distribution for postmerger data

Comparing the two distributions, shown by the solid and broken lines, respectively, we note the following two significant points First, the slope of the solid line is steeper than the slope of the broken line As discussed in Ijiri and Simon (1971), this slope measures the degree of concentration [10] Therefore, we can conclude that mergers and acquisitions in the 1948–69 period increased the degree of concentration [11]

The second important implication of table 1 is that the concavity of the distribution is increased as a result of mergers and acquisitions The vertical deviation between the curve (actual or estimated) and the line (theoretical) is much wider for the postmerger distribution than for the "mergerless" distribution [12]

IV. Summary and Conclusions

We have presented two different economic explanations for the concavity of the double-log firm-size distributions as observed in empirical data, one based on the autocorrelation of growth and the other based on mergers and acquisitions

---

[9] To some extent, the conclusions from tables 1 and 2 are affected by the assumption that all mergers and acquisitions have taken place at once A year-by-year analysis would indicate somewhat smaller differences among the rates of disappearance across the classes and among the rates of growth by asset reallocation across the classes

[10] It can be seen from eq (1) that if the rank, r, is increased to kr, the size, s, is reduced to s/k^{β}, indicating that the larger the β, the greater the relative size of a larger firm over a small firm when the ranks of the two firms are fixed It is clear from eq (2) that this β is the slope of the double-log distribution

[11] For the "mergerless" data, the size of the tenth firm is estimated by interpolation to be \$5 706 billion and the size of the eight hundred thirty-first firm is estimated to be \$0 064 billion from the data in column 5 of table 1 Connecting these points, β is estimated at 1 015 For postmerger data, the size of the tenth firm is \$6 146 billion and the size of the eight hundred thirty-first firm is \$0 050 billion from the *Fortune* data. Connecting these points, β is estimated to be 1.089, an increase of 7 percent in the slope as a result of mergers and acquisitions

[12] The greatest size variance (the ratio of actual over theoretical size) is approximately 1.6 for the "mergerless" data and 2.1 for the postmerger data.

Since both explanations are not only economically plausible but also supported by empirical data, it is likely that both factors have contributed to the concavity of the firm-size distribution Whether there are other contributing factors or whether these two factors are sufficient to eliminate most of the concavity is difficult to determine The degrees of autocorrelation reflected in γ would have to be determined independently from the observed curvature of the empirical distribution before we could answer this question At present, we have no such independent estimate of γ

However, the analysis in this paper does indicate the advantage of adjusting the Pareto curve for these additional economic factors instead of abandoning it by reason of the concavity in the observed distribution

From the standpoint of statistical curve fitting to empirical distributions, we may find other distributions to be superior to the modified Pareto curve However, it is not our sole objective to find a distribution that can best fit empirical data What we are after is a model that has sound economic support and justification and that can still give a reasonably good fit to the data The fact that the Pareto curve frequently approximates not only firm-size data but also a variety of economic and noneconomic skewed distributions is an important advantage in retaining it as a first approximation, instead of abandoning it

Appendix

Consider a firm which was created at epoch t as a result of getting one unit of sale The probability of getting a sale at epoch $(t + 1)$ is δ Its weight is discounted from 1 to γ at epoch $(t + 2)$ Hence, if the firm fails to receive the sale at epoch $(t + 1)$, the probability of a sale at epoch $(t + 2)$ is $\delta\gamma$ On the other hand, if the firm experiences a sale at epoch $(t + 1)$, its weight is changed to $(\gamma + 1)$, hence the probability of a sale at epoch $(t + 2)$ is $(\delta\gamma + \delta)$ Note that in either case the original sale contributes $\delta\gamma$ to the probability of a sale at epoch $(t + 2)$ regardless of whether there was a sale at epoch $(t + 1)$

In general, a sale at epoch t contributes to the probability of a sale at epoch $(t + k)$ by the amount $\delta\gamma^{k-1}$ Thus, if we focus only on the contribution by the *original* sale, this contribution decreases geometrically from δ for epoch $(t + 1)$, $\delta\gamma$ for epoch $(t + 2)$, $\delta\gamma^2$ for epoch $(t + 3)$, and so on Note that the events at each epoch (based only on the contribution to probability by the original sale) are mutually independent. Therefore, we have a sequence of independent Bernoulli trials where the probability of success decreases geometrically. The sum of sales generated $N_1 = X_1 + X_2 + \cdots +$, where X_i is 1 if a sale is generated at epoch $(t + i)$ and 0 otherwise, has the generating function

$$P_1(s) = [1 - \delta(1 - s)][1 - \delta\gamma(1 - s)][1 - \delta\gamma^2(1 - s)] \cdots \qquad (16)$$

Note that $P_1(1) = 1$. The subscript 1 in N_1 and P_1 is used to indicate that this expression involves only the sales generated from the original sales.

Hence, the derivative of $P_1(s)$ with respect to s is

$$
\begin{aligned}
P_1'(s) &= \frac{P_1(s)\delta}{1-\delta(1-s)} + \frac{P_1(s)\delta\gamma}{1-\delta\gamma(1-s)} + \frac{P_1(s)\delta\gamma^2}{1-\delta\gamma^2(1-s)} \\
&= P_1(s)\sum_{k=0}^{\infty}\frac{\delta\gamma^k}{1-\delta\gamma^k(1-s)} \qquad (17)
\end{aligned}
$$

The expected value of N, $E(N)$, is then given by

$$E(N_1) = P_1'(1) = P_1(1)\,\delta\sum_{k=0}^{\infty}\gamma^k = (1-\alpha)(1-\gamma)\frac{1}{1-\gamma} = 1-\alpha. \qquad (18)$$

The second derivative, $P_1''(s)$, is from (17)

$$
\begin{aligned}
P_1''(s) = P_1'(s)&\sum_{k=0}^{\infty}\frac{\delta\gamma^k}{1-\delta\gamma^k(1-s)} \\
&- P_1(s)\sum_{k=0}^{\infty}\frac{\delta^2\gamma^{2k}}{[1-\delta\gamma^k(1-s)]^2} \qquad (19)
\end{aligned}
$$

But from (17)

$$\sum_{k=0}^{\infty}\frac{\delta\gamma^k}{1-\delta\gamma^k(1-s)} = \frac{P_1'(s)}{P_1(s)} \qquad (20)$$

Substituting this in (19),

$$P_1''(s) = \frac{[P_1'(s)]^2}{P_1(s)} - P_1(s)\sum_{k=0}^{\infty}\frac{\delta^2\gamma^{2k}}{[1-\delta\gamma^k(1-s)]^2} \qquad (21)$$

Using the result in (18),

$$
\begin{aligned}
P_1''(1) &= (1-\alpha)^2 - \delta^2\sum_{k=0}^{\infty}\gamma^{2k} \\
&= (1-\alpha)^2 - (1-\alpha)^2(1-\gamma)^2\frac{1}{1-\gamma^2} = (1-\alpha)^2\frac{2\gamma}{1+\gamma} \qquad (21a)
\end{aligned}
$$

Now the variance of N_1, var (N_1), is given (see Feller 1968, p. 266) by

$$
\begin{aligned}
\text{var}\,(N_1) &= P_1''(1) + P_1'(1) - [P_1'(1)]^2 \\
&= (1-\alpha)^2\frac{2\gamma}{2+\gamma} + (1-\alpha) - (1-\alpha)^2 \\
&= (1-\alpha)\left[1 - \frac{(1-\alpha)(1-\gamma)}{1+\gamma}\right]. \qquad (22)
\end{aligned}
$$

Thus, var $(N_1) \to 1-\alpha$ as $\gamma \to 1$ and var $(N_1) \to \alpha(1-\alpha)$ as $\gamma \to 0$.

Note that if γ or α is close to 1, $\delta = (1-\alpha)(1-\gamma)$ is close to 0; therefore, applying the approximation (8) to (16), we obtain

$$\log P_1(s) = -(1-s)\delta(1+\gamma+\gamma^2 \quad) = -(1-s)(1-\alpha) \qquad (23)$$

or

$$P_1(s) = e^{-(1-\alpha)(1-s)}, \qquad (24)$$

which is a generating function of a Poisson distribution with mean $1 - \alpha$. However, regardless of the value of γ or α, the mean and the variance of N_1 given in (18) and (22) are exact and not approximations.

Now let us take into account the fact that the sales generated by the original sale also contribute to generating further sales. Thus, we have a branching process, each new sale having the potential of generating further sales. We may call the original sale a zero generation sale and a sale generated as a result of an nth generation sale an $(n + 1)$ generation sale.

Strictly speaking, the independence assumption needed for the application of theorems in branching processes is not satisfied here For example, at epoch $(t + k)$, the probability of the original sale generating another one is $\delta\gamma^{k-1}$ At the same time, if the original sale has generated another sale at epoch $(t + 1)$, the latter also has a probability of generating another sale at epoch $(t + k)$ equal to $\delta\gamma^{k-2}$ The two events are not independent It is not possible for the original sale and the first-generation sale both to generate sales, because only one unit of sales occurs at epoch $(t + k)$ However, by reducing the time interval between epochs and making a corresponding adjustment to γ, δ can be made as small as we want Hence, the probability of having two or more sales occurring in a given epoch may be considered to be negligible

Let N_k be the size of the kth-generation sales and P_k be the generating function of its probability distribution Since there is only one original sale, we set $N_0 = 1$. The generating function of N_1 has already been analyzed above

The kth-generation sales, N_k, may be divided into N_1 groups based on their ancestor in the first generation The number in each group has the same probability distribution as N_{k-1} Since they may be considered mutually independent approximately, the generating function P_k is given by a compound function (see Feller 1968, pp 295–96)

$$P_k(s) = P(P_{k-1}(s)) \tag{25}$$

From this, we obtain

$$P_k'(s) = P_1'(s)P_{k-1}'(s) = (P_1'(s))^k, \tag{26}$$

$$P_k''(s) = k(P_1'(s))^{k-1}P_1''(s) \tag{27}$$

Hence, $E(N_k)$ and var (N_k) are

$$E(N_k) = P_k'(1) = (1 - \alpha)^k, \tag{28}$$

$$\begin{aligned}
\text{var}\,(N_k) &= P_k''(1) + P_k'(1) - (P_k'(1))^2 \\
&= k(P_1'(1))^{k-1}P_1''(1) + (1 - \alpha)^k - (1 - \alpha)^{2k} \\
&= k(1 - \alpha)^{k-1}\left[(1 - \alpha)^2 - (1 - \alpha)^2\frac{1 - \gamma}{1 + \gamma}\right] \\
&\quad + (1 - \alpha)^k - (1 - \alpha)^{2k} \\
&= k(1 - \alpha)^{k+1}\frac{2\gamma}{1 + \gamma} + (1 - \alpha)^k - (1 - \alpha)^{2k}.
\end{aligned} \tag{29}$$

Now let M_k be

$$M_k = N_0 + N_1 + \cdots + N_k \tag{30}$$

and let

$$M = \lim_{k \to \infty} M_k,$$

namely, the ultimate size of a firm The generating function of M must satisfy

$$\rho(s) = sP(\rho(s)) \tag{31}$$

(Feller 1968, p 298) Hence,

$$\rho'(s) = P(\rho(s)) + sP'(\rho(s))\rho'(s), \tag{32}$$

$$\rho'(s) = P(\rho(s))/[1 - sP'(\rho(s))], \tag{33}$$

$$E(M) = \rho'(1) = \frac{1}{1 - (1 - \alpha)} = \frac{1}{\alpha} \tag{34}$$

Also from (32)

$$\rho''(s) = 2P'(\rho(s))\rho'(s) + sP''(\rho(s))(\rho'(s))^2 + sP'(\rho(s))\rho''(s), \tag{35}$$

$$\rho''(s) = \frac{2P'(\rho(s))\rho'(s) + sP''(\rho(s))(\rho'(s))^2}{1 - sP'(\rho(s))} \tag{36}$$

Hence,

$$\rho''(1) = \frac{1}{\alpha}\left[\frac{2(1-\alpha)}{\alpha} + \frac{(1-\alpha)^2}{\alpha^2}\,\frac{2\gamma}{1+\gamma}\right] \tag{37}$$

The variance of M is

$$\begin{aligned} \operatorname{var}(M) &= \rho''(1) + \rho'(1) - [\rho'(1)]^2 \\ &= \frac{1}{\alpha}\left[\frac{2(1-\alpha)}{\alpha} + \frac{(1-\alpha)^2}{\alpha^2}\,\frac{2\gamma}{1+\gamma} + 1 - \frac{1}{\alpha}\right] \\ &= \frac{1-\alpha}{\alpha^2} + \frac{(1-\alpha)^2}{\alpha^3}\,\frac{2\gamma}{1+\gamma} \end{aligned} \tag{38}$$

Thus, for $0 < \alpha < 1$, the growth process of a firm dies out with probability 1 since $E(M) = 1/\alpha < \infty$ Furthermore, we note that $E(M)$ depends only on α and is independent of γ In addition, var (M) is decreased as γ is decreased (namely, heavier discounting) However, since the total number of firms generated depends only on α and epoch t, the tail end of the firm-size distribution for the largest rank is fixed for given α and t and is independent of γ Thus, in order to reduce the variance of M, the sizes of above-average firms must be reduced and the sizes of below-average firms must be increased, thus creating more curvature in consistency with our earlier observations

References

Federal Trade Commission *Large Mergers in Manufacturing and Mining, 1948–1969* Statistical Report no 5 Washington Government Printing Office, February 1970

Feller, W *An Introduction to Probability Theory and Its Applications* Vol 1 3d ed. New York Wiley, 1968

Fortune "The 500 Largest U S Industrial Corporations," *Fortune*, May 1970. (*a*)

——— The Second 500 Largest U S Industrial Corporations," *Fortune*, June 1970 (*b*)

Ijiri, Y , and Simon, H A "Business Firm Growth and Size " *A.E R* 54 (March 1964) 77–89

——— "A Model of Business Firm Growth " *Econometrica* 35 (April 1967). 348–55

———. "Effects of Mergers and Acquisitions on Business Firm Concentration." *J.P.E.* 79 (March/April 1971): 314–22.

Simon, H. A. "On a Class of Skew Distribution Functions." *Biometrica* 52 (December 1955): 425–40.

Steindl, J. *Random Process and the Growth of Firms: A Study of the Pareto Law.* New York: Hafner, 1965

Soviet Growth in Absence of Centralized Planning: A Hypothetical Alternative

---

Micha Gisser and Paul Jonas

University of New Mexico

The study makes an attempt to investigate whether the Soviet Union would be richer or poorer if it had adopted a decentralized, agriculture-propelled, market-oriented economic policy The Soviet economy is divided into agricultural and nonagricultural sectors Each sector is approximated by a Cobb-Douglas production function Special output, labor, and capital indices are then used in order to compare the actual with the hypothetical growth of the Soviet economy The conclusion reached is that in absence of centralized planning, the growth of the USSR would have been the same or better

I. Introduction

Would the Soviet Union be richer or poorer if it had adopted a decentralized economic policy?

This question has been raised but never answered [1] The "middle-échelon" students of Soviet-type systems would argue that the question is trivial, since it is recognized that a centrally planned system is able to generate a process of speedy industrialization and rapid economic growth These views are summarized by Campbell "Explained in the simplest terms, the achievement has been to increase the level of Soviet economy at a pace scarcely ever achieved in human history . The Soviet system has made it possible to achieve unprecedented rates of growth" (Campbell 1960, p 29)

What are the main characteristics of this Soviet-type system that is able to generate the suggested successful economic performance? The works of Baykov (1947), Dobb (1948), Erlich (1950, 1955, 1960), Gerschenkron (1960), and others have revealed the various characteristics of a Soviet-type development policy. The most important prerequisites for a "classical" centrally planned development process are.

---

[1] Levine, for example, suggests that the examination of the possible course of the Soviet economy in the absence of central planning would be interesting. He feels, however, that the analytical difficulties involved "defy meaningful handling" (Levine 1968)

1 A dictatorship should take the factors of production, except labor, into state ownership, and coercion should be applied not only to the *ancien régime*, but also to every social stratum In other words, policy makers and activists should recognize that a new society based on central planning cannot be established without the extraordinary efforts and tears of the working class

2 Therefore, a charismatic leader should utilize the idea of creating a "golden age" to mobilize the labor force and should request or press for "voluntary ' and undercompensated achievements and high moral standards for workers and lower-échelon cadres At the same time, decision-making centers must tacitly accept the fact that there is no special economic-ideological feature in a Soviet-type system and that such an organization is subordinate to the same problems of scarcity and resource allocation as any other system

3 The agricultural sector should serve as a reservoir for capital and manpower for various industrial projects First, the peasantry should be neutralized, therefore, the latifundia of the landowners should be distributed among those "who till the soil " Peasant landownership, however, should not last long The strong, if not mythical, ties between the peasantry and the land should be broken Forced collectivization, compulsory deliveries, harsh administrative methods, and high taxes all provide excellent means to achieve this end, when the political power is firmly established These methods will induce an exodus of the active and young agricultural labor force from a sector that can be characterized by disguised unemployment Moreover, they will create capital accumulation from a segment of the economy where potential savings exceed actual savings

4 A goal-oriented command economy should now be organized, giving full economic power to a central planning board which orders the composition of the product mix in such a way that the output of capital goods must always exceed replacements A distorted price system, which makes inputs cheap and consumption expensive, will ease the way to achieve this goal

5 Finally, rigidly enforced development plans must be formulated Activists in various "transmission belts" and managers are expected to mobilize the workers for the fulfillment of targets Industrialization should start discontinuously, as a sudden great spurt, combined with the proliferation of every kind of heavy industry in a closed market virtually isolated from the outside world and considering this import substituting and unbalanced industrialization strategy as the most important means of expansion

To implement the outlined economic strategies,[2] the state should act as an Aztec-like god, demanding from the population more and more sacrifices and regarding personal consumption not as a goal but as an unfortunate input in the process of production. It is argued, however, that while the incurred social costs are painful, they are unavoidable in a process which, in a relatively short period, transforms a backward country into one that is just, equitable, and industrialized. The command model was considered a reliable generator of development, but recently the picture of accelerated economic growth in the Soviet Union, and generally in the Soviet bloc, has changed. Comments such as "the recent appearance of decline in output growth rates is now a commonplace observation among the students of Soviet economy" (Kaplan 1968) have become more and more frequent.[3]

The Soviet Union has arrived at a stage when economic stimuli are needed to create growth, and the planners have to realize that, in a complex economic milieu, scientific prognostication and effective stimulation of production through profitability should be introduced.

This recent trend has created a dramatic transformation in the camp of Sovietologists. What remains that is appealing in a command system, one may ask, if rapid economic growth can no longer be considered a success indicator?

A few years ago, the views of Jan Drewnowski (1961) were generally accepted. Drewnowski argued that the Lange-Lerner model of market socialism did not explain any reality because the suppositions are far away from the existing conditions. "Economists," he warned, "must not indulge in building castles in the air" (p. 342).

With the present shortages in the Soviet Union, which may originate from a phase when agriculture was exploited or, at the best, disregarded, more and more Sovietologists turn their attention to the examination of a Yugoslav-type model, or some variation of the old dream: socialism without coercion, socialism with a human face. In this latter framework, decision making is decentralized, and extensive use is made of the market mechanism and financial instruments which influence the direction and the coordination of the economy (Bergson 1967; Roberts 1968). Therefore, in the spirit of our time, we also will engage ourselves in "building castles in the air"

---

[2] The works of Erlich reveal that the person mainly responsible for this type of development strategy was Evgenii Preobrazhensky (1965, Nurkse 1957), a brilliant Russian theoretical economist and unsuccessful politician who died as one of the victims of Stalin's purges.

[3] Growth rates decline from about the early 1960s, and the concealment of the 1963 national income figure until mid-1964 gave ground for suspicion that a new growth trend had begun. The official Soviet estimate for this year was that the gross social product rose by 5 percent. The American governmental agency, the CIA, estimated the 1963 increase as 2.5 percent. Independent Western estimates ranged from zero (Warren Nutter) to the acceptance of the official Soviet figure (see Alec Nove, "2½ Per Cent and All That" [1964]).

We select for our suppositions the policy recommendations of Nikolai Bukharin, the spokesman of the opposition to the centrally planned course in the famous Soviet industrialization debate He rejected the idea that peasants would never provide resources for industrialization and, therefore, objected to the proposition that the government should forcefully collectivize agriculture and take resources by using coercion (Bukharin 1964)[4] Bukharin suggested an "agriculture-propelled" economic development, that is, that agriculture instead of heavy industry be emphasized and provided economic and technological help, and that a stream of savings originating from a prosperous peasantry should be plowed back into industrial expansion by the government He even proposed to the peasantry "*Enrichissez vous*"

We are aware of all the dangers and analytical difficulties associated with an examination concerned with the question "What would have happened in the USSR if, instead of a rigid Stalinist centrally planned course, Bukharin's flexible ideas had prevailed?" This problem looks, however, extremely challenging, and one ought not to shy away from the answer

For our examination, in accordance with Bukharin's position, we will make the following assumptions (1) among the factors of production, only capital is in state ownership, (2) the state adopts an agriculture-oriented policy in which the peasantry is not only protected from arbitrariness and requisitions, but is also induced to benefit from agrotechnical knowledge available in the West and is allowed to pursue various agricultural activities freely, and (3) the whole economy is cast in a decentralized New Economic Policy framework where the nationalized industry and the private agriculture are linked together by a semiprivate, flexible trading apparatus

II. The Actual Growth of Soviet Agricultural Sector

In order to examine the growth of the Soviet agricultural sector in the past, we construct a simple model with which we can separate the effects of labor, capital, and technology The Soviet economy is assumed to be governed by the Cobb-Douglas production function with neutral technological change as follows.

$$X = A(t)\ K^{\alpha}\cdot L^{1-\alpha}, \tag{1}$$

---

[4] Jasny seems to concur with Bukharin's views, when he states "The only conclusion that can be reached on the basis of the Soviet experience is that collectivization of agriculture should never be attempted by the methods resorted to in the USSR" (Jasny 1950)

TABLE 1
SHARE ESTIMATES FOR CAPITAL AND LABOR

	CAPITAL			LABOR		
	Low	Bergson	High	High	Bergson	Low
Agriculture	0.15	0.25	0.35	0.85	0.75	0.65
Nonagriculture	0.20	0.30	0.40	0.80	0.70	0.60

where X, A, K, and L denote agricultural output, technological change, capital, and labor input

The choice of the Cobb-Douglas production function is the best for the period to be considered here This is mainly due to its simplicity, which is particularly an asset when data are very crude Another type of production function was used by Weitzman (1970) Note, however, that Weitzman applied his production function for a recent time period which is different from the era investigated here Moreover, he was particularly interested in the effect of a low elasticity of substitution between labor and capital on the reduction in the Soviet economic growth in the late 1960s Following Solow (1957), we can apply a simple formula for measuring the contribution of the change in technology to the change in output This is sometimes called the neutral shift in technology

$$\dot{X}/X = \dot{A}/A + \alpha(/\dot{K}K) + (1 - \alpha)(L/L), \tag{2}$$

where the dot denotes a first-order derivative with respect to time, and $\dot{A}/A$ measures the neutral shift in the agricultural-production function. The reader should note that in fact $\dot{A}/A$ means more than just a neutral shift in the production function Griliches (1963) has demonstrated that most of the estimated rates of disembodied technical progress can be allocated to quality changes in the productivity of specific inputs For example, a rise in the level of schooling would be considered an improvement of human agents Since this is such an important point, an entire section will be devoted to it later

Kaplan indices (1968) are used as a measure of the percentage changes of outputs and inputs

A more difficult problem is to obtain a rough estimate of the coefficients of capital and labor in the Cobb-Douglas production function Based on the findings of Bergson (1963), we adopt the following share estimates for capital and labor in the Soviet Union (see table 1)

The Kaplan indices are available for capital (excluding livestock), sown area, and man-years The omission of livestock introduces an error into the data. It is hoped that this error is the smallest error in the data. Based on Bergson (1963), it was assumed that nonland weighs 17.5 percent and

TABLE 2

Contribution of Technology Change to Growth of Soviet Agriculture

	1928–40	1951–54	1954–58	1958–63
Output	1.60	5.60	9.20	−0.20
Capital	3.20	4.90	6.10	3.60
Man-Years	1.20	−0.90	−0.70	0.80
Technology *				
Capital low	0.10	5.63	8.88	−1.42
Bergson	−0.10	5.05	8.20	−1.70
Capital high	−0.30	4.47	7.52	−1.98

Note.—Based on Kaplan indices (percentage change per annum).

* A/A, expressed as a percentage (see eq. [2]).

land 82.5 percent in total agricultural capital.[5] In table 2 we show the Kaplan indices for outputs and inputs and our estimate of the contribution of technology.

It can be seen that the actual productivity trend in agriculture in the period of 1928–40 was stagnant. During this period the ownership and control of the bulk of productive resources were gradually transferred to the state, several prices existed and were fixed more or less arbitrarily. The government extracted from agriculture a significant part of its value added by means of the regimen of compulsory deliveries to state procurement organizations at low fixed prices.

In the period starting from 1951, actual agricultural productivity in the Soviet Union shows a "great spurt" characteristic, and this culminates in the year 1958 and declines after that. The reason for this upsurge is that Soviet planners realized the lagging performance of agriculture and turned their attention to the problems of agricultural productivity. The government, during this period, subsidized to some extent the prices of some selected agricultural products, such as fiber and oil crops, and also gave help in the form of controlled prices for fertilizers, fuel, feed, and building materials which were lower than retail prices applicable to private use. Another subsidy, during the upsurge, consisted of direct grants to machine-tractor stations and state farms to cover their operating losses.

III. The Actual Growth of Soviet Nonagricultural Sector

In order to examine the growth of the nonagricultural sector, we use similar procedures. Let the nonagricultural sector be summarized by the following Cobb-Douglas production function

$$Y = B(t)\cdot K_{(t)}^{\beta} L_{(t)}^{1-\beta}, \tag{3}$$

---

[5] See Appendix A.

TABLE 3

CONTRIBUTION OF TECHNOLOGY CHANGE TO GROWTH OF SOVIET NONAGRICULTURAL SECTOR

	1928–40	1951–54	1954–58	1958–63
Output	8 10	7 00	6 50	5 90
Capital	14 10	9 20	10 10	11 30
Man-years	8 50	3 50	1 10	1 80
Technology *				
Capital low	−1 52	2 36	3 60	2 20
Bergson	−2 08	1 79	2 70	1 25
Capital high	−2 64	1 22	1 80	0 30

NOTE —Based on Kaplan indices (percentage change per annum)

* $\dot{B}/B$, expressed as a percentage (see eq [4])

where Y, B, $K_{(1)}$, and $L_{(1)}$ denote nonagricultural output, technological change, capital, and labor input Again, following Solow (1957), we obtain the measurement of the change in technology as follows

$$\dot{Y}/Y = \dot{B}/B + \beta\dot{K}_{(1)}/K_{(1)} + (1 - \beta)\ \dot{L}_{(1)}/L_{(1)} \qquad (4)$$

The estimates of the capital and labor shares based on Bergson are $\beta = 0\ 30$, $1 - \beta = 0\ 70$

Using Kaplan indices, we obtain the following measurements of the neutral shift in the aggregate nonagricultural production function

It is interesting to compare the results of tables 2 and 3 The percentage shifts in the productivity move roughly in the same direction in both agriculture and in the nonagricultural sector "Technology" in the nonagricultural sector between 1928 and 1940 declines, and at the same time this period shows an extremely rapid growth rate The decline of "technology" is counteracted by huge capital investments and by changing dramatically the capital-labor ratio This period kept Soviet agriculture stagnant, but produced in the still partly underutilized industrial sector a high growth rate The rapid growth in the nonagricultural sector declines a little bit after the reconstruction period, but the output in the agricultural sector increases significantly After Stalin's death, in the period of "general thaw," the slowdown in nonagricultural output rate begins, which is, interestingly, coupled with a dramatic rate of growth in the agricultural sector In the last phase of 1958–63 the direction is the same in the agricultural and nonagricultural sector, and they both contribute to the retardation in Soviet growth

It is possible to propose a hypothesis which would link certain fluctuations in the growth rates and productivity shifts with the changing institutional system in the Soviet Union The national purpose, which was loosely defined as "maximization of industrialization," may produce periods of investment booms that exceed the capacity of the economy, and these must be followed by periods of consolidation. We may also conclude that the fluctuations in the nonagricultural and especially in the agricultural output in the Soviet Union around a general trend line seem pronounced enough to constitute an addition to the evidence of economic cycles in Soviet-type economies A comparison with the behavior of other output series, especially those relating to investment in some East Central European Soviet-type economies, shows a marked similarity (Staller 1964)

It can be supposed that the recent retardation in Soviet growth can be attributed to long-run cycles in the Soviet agricultural-production function a period in which because of political reasons the agricultural sector does not absorb new technologies from abroad is followed by a period of political relaxation, in which the agricultural sector attempts to absorb the technologies it failed to adopt during the long period of political rigidity and make use of accumulated improvement in inputs which was not exploited due to an absence of economic freedom

IV. Disembodied Technical Progress

It helps if, following Griliches (1963), one decomposes the contribution of technology in agriculture into (*a*) improvement in the quality of inputs, (*b*) elimination of disequilibria, and (*c*) selecting the optimal scale. As far as improvement in the quality of human agents is concerned, one can conjecture that more schooling contributes more to agricultural productivity when the educated farmer is free to apply his education in adopting new techniques, better marketing procedures, and more effective management There was a considerable drive to increase the level of schooling in the USSR during 1928–40 It is interesting to note that agricultural production shows no response to this increased level of education in other words, the disembodied technological progress is not detectable What was the reason for this phenomenon? Was the centrally planned structure where decisions were made at a higher level and managers followed instructions step by step so rigid that the latter could not use their newly acquired skills? Did the abnormally depressed agricultural price system give too little incentive for the leaders of cooperatives to improve efficiency? Is it true that the most important gain from education was that it enabled agricultural workers to avoid regulations? The "great spurt" of 1951–58 probably shows that due to more favorable prices and relaxation of economic controls, agricultural employees could take advantage of the previous improvement in inputs, such as more schooling. For example, the rising level of schooling and improving machinery was helpless during the period prior to 1951 because of the known fact that the Soviets failed to develop or adopt hybrid seeds and chemical fertilizers We can also conjecture that during that early period in Soviet development, the resistance of peasants to the collectivization drive overwhelmed the rise in their level of schooling, as well as improvements in nonhuman inputs. Thus, findings in table 2 probably indicate that the process of collectivization not so much eliminated disembodied technical progress as it did concentrate it in a relatively short time period The reader should also note that for the period after 1950 there was a gradual change in the sex and age composition of the farm labor force The rural male population suffered heavy losses during World War II, and by 1950 agriculture was heavily dependent upon female labor The proportion of males in the labor force increased year by year thereafter, and this change in the sex ratio should have contributed to output over and above the changes in man-years of employment and schooling This leads us to the next section which boils down to assuming that had the Soviets not collectivized their agricultural sector, disembodied technical progress would have probably been spread more evenly over time In particular, the agricultural sector would have enjoyed a moderate increase in disembodied technology during the period 1928–40

V. The Hypothetical Alternative

In accordance with our assumption (2), we suppose that the Soviet Union had not collectivized its agricultural sector Rather, the Soviet Union allowed farmers to buy inputs and sell agricultural output in a free market in which prices are either determined according to laws of supply and demand, or, alternatively, administered by the government, but not in a manner which is unfavorable to farmers We assume also that the Soviet Union encouraged her private farmers to adopt all the new agricultural techniques that were made available in the United States and other Western countries Is it safe to assume that, under these assumptions, the Soviet agricultural sector could have benefited from disembodied technologies at a rate similar to the rate that had prevailed in the United States? Note that this is not an implied assumption, this is a historical question

If we accept the proposition that in absence of collectivization a free Soviet farm sector would have enjoyed a more evenly distributed disembodied technical change over time, we can take a further step and assume that such progress of disembodied technology would have followed the Western pattern over time Following Schultz (1953), we can accept the estimate of 1 percent increase in agricultural output per unit of input during the period 1928–40 The relevant rate for the post–World War II period for the United States is 2 percent (Heady 1967). Thus, the agricultural-production function would have shifted hypothetically at an annual rate of 1 percent from 1928 to 1940, and at a rate of 2 percent in the postwar period.

these must be followed by periods of consolidation. We may also conclude that the fluctuations in the nonagricultural and especially in the agricultural output in the Soviet Union around a general trend line seem pronounced enough to constitute an addition to the evidence of economic cycles in Soviet-type economies A comparison with the behavior of other output series, especially those relating to investment in some East Central European Soviet-type economies, shows a marked similarity (Staller 1964)

It can be supposed that the recent retardation in Soviet growth can be attributed to long-run cycles in the Soviet agricultural-production function: a period in which because of political reasons the agricultural sector does not absorb new technologies from abroad is followed by a period of political relaxation, in which the agricultural sector attempts to absorb the technologies it failed to adopt during the long period of political rigidity and make use of accumulated improvement in inputs which was not exploited due to an absence of economic freedom

IV. Disembodied Technical Progress

It helps if, following Griliches (1963), one decomposes the contribution of technology in agriculture into (*a*) improvement in the quality of inputs, (*b*) elimination of disequilibria, and (*c*) selecting the optimal scale. As far as improvement in the quality of human agents is concerned, one can conjecture that more schooling contributes more to agricultural productivity when the educated farmer is free to apply his education in adopting new techniques, better marketing procedures, and more effective management There was a considerable drive to increase the level of schooling in the USSR during 1928–40 It is interesting to note that agricultural production shows no response to this increased level of education: in other words, the disembodied technological progress is not detectable What was the reason for this phenomenon? Was the centrally planned structure where decisions were made at a higher level and managers followed instructions step by step so rigid that the latter could not use their newly acquired skills? Did the abnormally depressed agricultural price system give too little incentive for the leaders of cooperatives to improve efficiency? Is it true that the most important gain from education was that it enabled agricultural workers to avoid regulations? The "great spurt" of 1951–58 probably shows that due to more favorable prices and relaxation of economic controls, agricultural employees could take advantage of the previous improvement in inputs, such as more schooling For example, the rising level of schooling and improving machinery was helpless during the period prior to 1951 because of the known fact that the Soviets failed to develop or adopt hybrid seeds and chemical fertilizers. We can also conjecture that during that early period in Soviet development, the resistance of peasants to the collectivization drive overwhelmed the rise in their level of schooling, as well as improvements in nonhuman inputs Thus, findings in table 2 probably indicate that the process of collectivization not so much eliminated disembodied technical progress as it did concentrate it in a relatively short time period The reader should also note that for the period after 1950 there was a gradual change in the sex and age composition of the farm labor force The rural male population suffered heavy losses during World War II, and by 1950 agriculture was heavily dependent upon female labor The proportion of males in the labor force increased year by year thereafter, and this change in the sex ratio should have contributed to output over and above the changes in man-years of employment and schooling This leads us to the next section which boils down to assuming that had the Soviets not collectivized their agricultural sector, disembodied technical progress would have probably been spread more evenly over time In particular, the agricultural sector would have enjoyed a moderate increase in disembodied technology during the period 1928–40

V. The Hypothetical Alternative

In accordance with our assumption (2), we suppose that the Soviet Union had not collectivized its agricultural sector Rather, the Soviet Union allowed farmers to buy inputs and sell agricultural output in a free market in which prices are either determined according to laws of supply and demand, or, alternatively, administered by the government, but not in a manner which is unfavorable to farmers We assume also that the Soviet Union encouraged her private farmers to adopt all the new agricultural techniques that were made available in the United States and other Western countries Is it safe to assume that, under these assumptions, the Soviet agricultural sector could have benefited from disembodied technologies at a rate similar to the rate that had prevailed in the United States? Note that this is not an implied assumption, this is a historical question

If we accept the proposition that in absence of collectivization a free Soviet farm sector would have enjoyed a more evenly distributed disembodied technical change over time, we can take a further step and assume that such progress of disembodied technology would have followed the Western pattern over time Following Schultz (1953), we can accept the estimate of 1 percent increase in agricultural output per unit of input during the period 1928–40 The relevant rate for the post-World War II period for the United States is 2 percent (Heady 1967). Thus, the agricultural-production function would have shifted hypothetically at an annual rate of 1 percent from 1928 to 1940, and at a rate of 2 percent in the postwar period

We admit that borrowing the productivity growth rates from Schultz's and Heady's work on the United States overlooks the fact that these growth rates are not just a function of the specific time space in which they occurred, but also are related to the factor proportions and other practical circumstances obtained in the United States in those years. One might suggest that if a comparison of this kind is tried at all, it should be with some period in the United States history when the country's economic structure was somewhat more similar to the facts of Russia in the 1930s and 1950s. And yet, if one takes into consideration the fact that the Soviet government embarked on a process of mass education of the Russian peasants, and also that the Soviets did not have to invest in agricultural technologies, but only to borrow available technologies from the West, our assumption looks more attractive. Notice finally that we are not adding the change in disembodied agricultural technology in the United States to the Soviet change. We are replacing Soviet patterns by the Western pattern. For example, table 2 shows that agricultural output grew at an annual rate of 9.2 percent during 1954–58. Using the estimate based on Bergson, disembodied technology increased at a rate of 8.20 percent. This should be replaced by a 2 percent change as indicated above. Accordingly, the hypothetical growth rate of agricultural output should be 3.00 percent ($3.00 = 9.20 - 8.20 + 2.00$).

According to Eason (1963), the Soviet agricultural labor force amounted to 84.9 and 61.2 percent of the total labor force in 1926 and 1939, respectively. Thus, it is safe to assume that on the average, during that time period, 1 percent released from the agricultural labor force would have amounted to 2.5 percent increase in the nonagricultural labor force. Replacing 0.10, −0.10, or −0.30 percent change (table 2) by 1 percent would have resulted in a change of 0.9, 1.1, or 1.3 percent in disembodied technology. Since the agricultural share is assumed to be 0.85, 0.75, or 0.65, respectively, adding the above percentage changes and maintaining the same growth rate of agricultural output would have released more labor to the nonagricultural sector. As explained in Appendix B, this would have raised the annual growth rate of output in the nonagricultural sector from 8.1 to 10.22, 10.67, or 11.10 percent, respectively, for the period of 1928–40.

Note that the assumption that the agricultural output target for the year 1940 had not changed is somewhat arbitrary. Given our hypothetical model, this assumption allows a higher rate of farm out-migration which would have been favorable to the industrial sector. In other words, the process of farm out-migration is a channel through which progress in the agricultural sector can be "exported" to the industrial sector. Had we adopted for our hypothetical model an agricultural output target for the year 1940 higher than the actual target, less progress would have been transferred from the agricultural to the industrial sector. Nevertheless,

TABLE 4
Actual and Hypothetical Growth Rates of Output in the Two Sectors, USSR (Percentage Change per Annum)

	1928–40	1951–54	1954–58	1958–63
Nonagricultural				
Actual (Kaplan)	8.10	7.00	6.50	5.90
Hypothetical				
Low capital	10.22	7.00	6.50	5.90
Bergson	10.67	7.00	6.50	5.90
High capital	11.10	7.00	6.50	5.90
Agricultural				
Actual (Kaplan)	1.60	5.60	9.20	−0.20
Hypothetical				
Low capital	1.60	1.97	2.32	3.22
Bergson	1.60	2.55	3.00	3.50
High capital	1.60	3.13	3.68	3.78

Note.—For detailed explanation, see Appendix B.

TABLE 5
Actual and Hypothetical Growth Indices, USSR

	1928	1940	1951	1954	1958	1963
Nonagricultural						
Actual	100	254.63	272.45	333.77	429.38	571.90
Hypothetical						
Low capital	100	321.46	343.96	421.36	542.07	721.99
Bergson	100	337.57	361.19	442.48	569.23	758.17
High capital	100	353.65	378.40	463.56	596.35	794.29
Agricultural						
Actual	100	120.98	127.76	150.45	213.93	211.80
Hypothetical						
Low capital	100	120.98	123.37	130.80	143.37	167.99
Bergson	100	120.98	124.07	133.80	150.60	178.86
High capital	100	120.98	124.77	136.86	158.14	190.38

according to our assumption, the nonagricultural sector could have benefited from an additional 2.65, 3.67, or 5.00 percent growth of its labor force for 1928–40 (see Appendix B). Given a share of labor of 0.80, 0.70, or 0.60, this could have meant an additional 2.12, 2.57, or 3.00 percent on top of the 8.1 percent actually achieved. This means that hypothetical growth rates of outputs in the USSR could have been as shown in table 4.

Based on table 4, and assuming for the sake of argument that output levels in 1950 reached their previous 1940 level, we construct in table 5 indices which compare our hypothetical path of growth with the actual one.

In what follows we shall adhere to the estimates based on Bergson. However, the reader can check the results for low and high capital estimates as they are shown in table 5.

Table 5 shows that if our hypothetical alternative had been followed, in 1940 nonagricultural production in the USSR could have been 33 percent higher than the actual level. This would have been achieved without any sacrifice, social or economic, on the part of the agricultural sector.[6]

In 1963, the trade-off would have been slightly less favorable: nonagricultural output would have still been 33 percent higher than the actual level, but agricultural output would have had to be 14 percent less than it actually was.

The reader should not be puzzled by the fact that under our hypothesis the agricultural sector ends up a "loser" in 1963. This results from our assumption that the agricultural output target for the year 1940 had not changed under our hypothesis. Recall that because of this assumption the agricultural sector could have released 1.47 percent of its labor force, which meant 3.67 percent extra growth for the industrial labor force. In table 4, this labor transfer, while leaving the rate of growth of the hypothetical agricultural output unchanged at 1.6 percent, increases the rate of growth of the industrial sector from 8.1 to 10.67 percent for the period 1928–40.

Sensitivity Analysis (Variation in Assumptions)

Some sensitivity analysis has been used throughout the analysis. In applying the Cobb-Douglas production function, we used three estimates for the coefficient of capital: low, "Bergson," and high. But there are some other moderate variations in assumptions that can be made in order to test the validity of our model. We now consider some of the problems that arise in conjunction with sensitivity analysis. We admit that some of these problems cannot be solved, while other problems can be subjected to a meaningful test. Consider the following:

a) ***Would the various Kaplan indices be valid under our hypothetical assumptions?***—We can give a reasonable answer only when the labor input is considered. We believe the labor-input index would have remained the same, first due to the fact that population would have grown at the same rate over time. Second, we believe that labor-force participation rate would have remained about the same, mainly because incentives to join the labor force would have replaced coercion.

What about capital indices? In this case we can only speculate. Would saving be lower under our hypothetical alternative? Even if domestic saving would have been lower, it is very likely that borrowing capital in "capitalistic" markets abroad would have become easier, thus making up for the reduction in domestic saving.

---

[6] Note that actual agricultural output in 1958 and 1963 were "abnormal," the first too high and the second too low because of climatic conditions. Accordingly, if one uses a 3-year average output level centered on 1958, the adjusted actual output for 1958 in table 5 would be 196.8, using the same process gives an adjusted output for 1963 of 222.4.

b) *It is very likely that prices received by farmers would have been much higher under our hypothetical alternative compared with actual prices.*—This gives rise to a problem which is closely related to the following:

c) *It is not at all clear that the Russian industrial sector was ready to absorb more labor released by the agricultural sector.*—Let us now discuss (*b*) and (*c*) and test a variation in our assumption concerning the rate of farm out-migration. If agriculture and the rest of the economy had been linked together by a semiprivate, flexible trading apparatus, there would have been fewer persons transferring from agriculture to industry during 1928–40 than under our hypothetical alternative, or compared even with what in fact occurred. The prices received by farmers would have been much higher than the nominal prices at which much of the farm output was transferred during the thirties.

Next, there are doubts that industrial employment could have grown at a significantly higher rate from 1928 to 1940 than it did without resulting in an even greater rate of negative disembodied technology. This results from assuming that no technological change, either positive or negative, would have been associated with the transfer of additional labor, while during the 1928–40 period the data in table 3, using Bergson's coefficients, implied that the actual annual growth in inputs "overexplained" the growth of output by an annual rate of slightly more than 2 percent. Thus, it is not clear how further straining of organizational capacity could result in the ability to expand size further as required to absorb an industrial labor force growing at 11.2 percent annually instead of 8.5 percent.

To test this variation in our assumptions, let us revise our hypothetical alternative. Assume that under the semiprivate, flexible trading apparatus, the rate of farm out-migration would have been the same as the actual rate. This means a reduction of the farm labor force from slightly over 80 percent in 1928 to 60 percent in 1940 and 40 percent in 1963. It is reasonable to assume that proper taxes and subsidies coupled to a free-market mechanism would have achieved such a goal. Under such assumptions, the hypothetical growth rates in the agricultural sector for the period 1928–40 would have been 2.50, 2.70, and 2.90, respectively. (For example, 2.50 = 3.20 × 0.15 + 1.20 × 0.85 + 1.00.) These rates would replace 1.60 percent in table 4. Resulting from such an adjustment, table 6 would replace table 5, as follows. Here, the agricultural sector would have ended up at 203.53, which is slightly less than the actual 211.80. What it means is that, even with this modification of assumptions, under a free market the Soviet Economy would have performed as well as it actually did.

TABLE 6
ACTUAL AND HYPOTHETICAL GROWTH INDICES, USSR, ASSUMING NO CHANGE IN FARM OUT-MIGRATION RATES

	1928	1940	1951	1954	1958	1963
1 Nonagricultural						
Actual = Hypothetical	100	254.63	272.45	333.77	429.38	571.90
2 Agricultural.						
Actual	100	120.98	127.76	150.45	213.93	211.80
Hypothetical						
Low capital	100	134.49	137.14	145.40	159.37	186.73
Bergson	100	137.67	141.18	152.26	171.37	203.53
High capital	100	140.92	145.33	159.40	184.19	221.73

VI. Conclusion

The result of our analysis is that industrialization without the "super-industrializers" could have occurred at the same rate or even a more impressive rate than actually happened. The computed annual growth rate in the nonagricultural sector could have been 2.57 percent on top of the 8.1 percent during the period 1928–40. This could have been achieved at no social and economic cost to the agricultural sector until 1940, and at a small economic cost (but without the incurred social cost) to the agricultural sector thereafter.

Our results suggest that there was the same or even greater growth possibility in the approach of those who, in the "Great Industrialization Debate," argued that a market relationship between a free agricultural sector and a state-owned industrial sector could generate savings and investments in the former and could reduce costs and establish profitability in the latter. In this way more capacity could have been created in industry without sacrificing agriculture and living standards. The computations show that this idea was far from being utopian. The acceptance of the Preobrazhensky-Stalin path led to unnecessary sufferings on the part of the Soviet population and misallocation of resources.

We admit that our model gives rise to some difficulties that were pointed out in the last section.

First, if agriculture and the rest of the economy had been linked together by a "semiprivate, flexible trading apparatus," there would have been fewer persons transferring from agriculture to industry during 1928–40 than we have suggested. The prices received by the farmers would have been higher than the nominal prices at which much of the farm output was transferred during the thirties. Moreover, as indicated in the last section, we have doubts about the capacity of the Soviet industrial sector to absorb more labor than it actually did during 1928–40. Adjusting our assumptions accordingly yielded table 6. All these findings are summarized in table 7 as follows. Thus, even if we assume no change in the rate of farm

TABLE 7
SUMMARY OF FINDINGS INDICES

	1928	1940	1963
1 Nonagricultural			
Actual	100	254.63	571.90
Hypothetical, assuming an accelerated rate of farm out-migration	100	337.57	758.17
Hypothetical, no change in the rate of farm out-migration	100	254.63	571.90
2 Agricultural:			
Actual	100	120.98	211.80
Hypothetical, assuming an accelerated rate of farm out-migration	100	120.98	178.86
Hypothetical, assuming no change in the rate of farm out-migration	100	137.67	203.53

NOTE.—Capital coefficients are according to Bergson.

out-migration, we have shown that the Soviet Union could have achieved the same economic miracles under a system which is closer to a free-market mechanism. We conclude that the heavy price in human life, in human suffering, and in lack of freedom that many Soviet citizens had to pay for adopting the Preobrazhensky-Stalin system was not needed. It seems that Bukharin was right after all.

Appendix A

The following (table A1) is a revised version of Bergson (1963) of the shares of capital and labor:

TABLE A1
BERGSON'S SECTOR ESTIMATES

	BOTH SECTORS TOGETHER (Billions of Rubles)	AGRICULTURAL SECTOR		NONAGRICULTURAL SECTOR	
		Billions of Rubles	% Share	Billions of Rubles	% Share
Set A					
Employment	179.9	110.1*	75	69.8	64
Capital	46.2	6.9†		39.3	36
Farm land	22.4	22.4	25		
Livestock	7.0	7.0			
Total	255.5	146.4	100	109.1	100
Set B					
Employment	179.9	110.1*	80	69.8	82
Capital	18.5	2.8†		15.7	18
Farm land	22.4	22.4	20		
Livestock	2.8	2.8			
Total	223.6	138.1	100	85.5	100
Set C					
Employment	179.9	110.1*	70	69.8	64
Capital	46.2	6.9†		39.3	36
Farm land	33.6	33.6	30		
Livestock	7.0	7.0			
Total	266.7	157.6	100	109.1	100

* Agricultural labor force accounted for 61.2 percent of the entire labor force in 1939. This is based on findings of Eason (1963).
† We assumed that 15 percent of the item capital was agricultural.

Appendix B

Based on Eason (1963), during the period 1928–40, 1 percent of labor released by the agricultural sector would have contributed 2.5 percent increase in the labor force of the nonagricultural sector. Based on Bergson (1963), the agricultural labor share was 0.75 and the nonagricultural labor share was 0.70. Since the data are crude, we considered a range of shares, rather than a point estimate. The hypothetical growth rate of the nonagricultural sector for the period 1928–40 was calculated, as shown in table 3.

The hypothetical growth rate of agricultural output for 1928–40 is unchanged. This is true because the agricultural production goal would have remained the same for the year 1940 under our hypothesis. But, replacing the change in disembodied technology amounting to 0.10, –0.10, and –0.30 by 1 percent change we obtained 0.9, 1.1, or 1.3 alternative hypothetical disembodied technological changes. This would have enabled the agricultural sector to release 1.06, 1.47, or 2.00 percent per annum of its labor force during 1928–40.

TABLE B1

COMPUTATION OF LABOR AND OUTPUT CHANGES

	Agricultural Labor Share $(1 - \alpha)$	Labor Released from Agricultural Labor (% per Year)	Additional Change in Nonagricultural Labor (% per Year)	Nonagricultural Labor Share $(1 - \beta)$	Additional Change in Nonagricultural Output (% per Year)	Hypothetical Change in Nonagricultural Output (% per Year)
Low capital	0.85	0.9/0.85 = 1.06	1.06 × 2.5 = 2.65	0.80	2.65 × 0.80 = 2.12	8.1 + 2.12 = 10.22
Bergson	0.75	1.1/0.75 = 1.47	1.47 × 2.5 = 3.67	0.70	3.67 × 0.70 = 2.57	8.1 + 2.57 = 10.67
High capital	0.65	1.3/0.65 = 2.00	2.00 × 2.5 = 5.00	0.60	5.00 × 0.60 = 3.00	8.1 + 3.00 = 11.10

Next, the annual growth rates of agricultural output for the postwar periods were obtained by replacing the actual percentage changes in disembodied technology with our 2 percent hypothetical change, as follows.

1951–54

$5.6 - 5.63 + 2 = 1.97$

$5.6 - 5.05 + 2 = 2.55$

$5.6 - 4.47 + 2 = 3.13$

1954–58

$9.2 - 8.88 + 2 = 2.32$

$9.2 - 8.20 + 2 = 3.00$

$9.2 - 7.52 + 2 = 3.68$

1958–63

$-0.2 - (-1.42) + 2 = 3.22$

$-0.2 - (-1.70) + 2 = 3.50$

$-0.2 - (-1.98) + 2 = 3.78$

References

Baykov, Alexander. *The Development of the Soviet Economic System.* Cambridge, Mass.: Harvard Univ. Press, 1947.

Bergson, Abram. "National Income." In *Economic Trends in the Soviet Union,* edited by A. Bergson and S. Kuznetz. Cambridge, Mass.: Harvard Univ. Press, 1963.

———. "Market Socialism Revisited." *J.P.E.* 75, no. 4, suppl. (October 1967): 655–76.

Bukharin, N. I. "Notes of an Economist at the Beginning of a New Economic Year." Reprinted in *Foundations of Soviet Strategy for Economic Growth: Selected Soviet Essays,* edited by Nicolas Spulber. Bloomington: Indiana Univ. Press, 1964.

Campbell, Robert W. *Soviet Economic Power.* Cambridge, Mass.: Riverside Press, 1960.

Dobb, Maurice. *Soviet Economic Development since 1917.* New York: n.p., 1948.

Drewnowski, Jan. "The Economic Theory of Socialism: A Suggestion for Reconsideration." *J.P.E.* 69, no. 4 (August 1961): 341–54.

Eason, Warren W. "Labor Force." In *Economic Trends in The Soviet Union,* edited by A. Bergson and S. Kuznetz. Cambridge, Mass.: Harvard Univ. Press, 1963.

Erlich, Alexander. "Preobrazhenskii and the Economics of Soviet Industrialization." *Q.J.E.* 69 (February 1950): 57–88.

———. "Stalin on Economic Development." In *Continuity and Change in Russian and Soviet Thought,* edited by E. Simmons. Cambridge, Mass.: Harvard Univ. Press, 1955.

———. *The Soviet Industrialization Debate,* 1924–1929. Cambridge, Mass.: Harvard Univ. Press, 1960.

Gerschenkron, Alexander. "Problems and Patterns of Russian Economic Development." In *Transformation of Russian Society,* edited by Cyril Black. Cambridge, Mass.: Harvard Univ. Press, 1960.

Griliches, Zvi. "The Sources of Measured Productivity Growth: United States Agriculture." *J.P.E.* 171, no. 4 (August 1963): 331–46.

Heady, Earl O. *A Primer on Food Agriculture and Public Policy.* New York: Random House, 1967.

Jasny, Naum. *The Socialized Agriculture of the U.S.S.R.: Plans and Performances.* Stanford, Calif.: Stanford Univ. Press, 1950.

Kaplan, Norman M "Retardation in Soviet Growth." *Rev Econ and Statis.* 50, no 3 (August 1968) 293–303

Levine, Herbert S "The Role of Central Planning." In *New Currents in Soviet-Type Economics*, edited by George R Feiwel Scranton, Pa. International Textbook Co, 1968.

Nove, Alec. "2½ Per Cent and All That" *Soviet Studies* 16, no. 4 (July 1964) 28–54

Nurkse, Ragnar *Problems of Capital Formation in Underdeveloped Countries* New York Oxford Univ Press, 1957.

Preobrazhensky, Evgeny *The New Economics* Oxford Clarendon Press, 1965

Roberts, Paul Craig "Drewnowski's Economic Theory of Socialism" *J P E* 76, no 4 (July/August 1968) 645–50

Schultz, Theodore W *The Economic Organization of Agriculture*, New York McGraw-Hill, 1953

Solow, Robert M "Technical Change and the Aggregate Production Function" *Rev Econ and Statis* 39, no 3 (August 1957) 312–20

Staller, George J. "Fluctuations in Economic Activity Planned and Free Market Economies, 1950–1960" *A.E R* 39, no 3 (June 1964) 312–20

Weitzman, Martin L "Soviet Postwar Economic Growth and Capital-Labor Substitution" *A.E R* 60, no 4 (September 1970) 676–92

Ghetto Employment and the Model Cities Program

Bennett Harrison

Massachusetts Institute of Technology

In 1971, the Model Cities Program subsidized 25,000 jobs in municipal and nonprofit agencies Both the original legislation and subsequent administrative rulings require participating cities to give preference in hiring to residents of the target areas—the urban ghetto. Yet, by the spring of 1971, fewer than half the positions were held by ghetto residents Moreover, nonresidents were being paid significantly higher salaries than residents. This paper attempts to identify the determinants of intercity variation in "resident participation rates" and the causes of the resident/nonresident differences in earnings It was found that cities with highly structured bureaucracies and/or strong mayors allocate a larger share of Model Cities jobs to "outsiders" than is the case for other cities. On the other hand, real or potential "black power" (as measured, for example, by whether or not the city experienced a major civil disorder in 1967) increases ghetto participation in the Model Cities Program. Analysis of more than 14,000 individual employee records shows that neither occupational "crowding" nor lack of educational credentials sufficiently explains the salary differences between ghetto and non-ghetto Model Cities workers

Introduction

During the 1960s, the federal government launched a panoply of anti-poverty programs designed to ameliorate conditions in the ghettos of American cities Government planners, academic researchers, and representatives of the organized poor themselves have focused their

---

The author is associate professor of economics and urban studies, Department of Urban Studies and Planning, M I T This paper was written under a grant from the U.S Office of Economic Opportunity to the Maryland Project on the Economics of Discrimination, and was delivered to the 1972 annual meetings of the American Political Science Association Data and guidance in their use were provided by the U.S. Department of Housing and Urban Development and by the National Civil Service League Bradley Schiller, Charles Levine, and Frank Thompson contributed helpful criticism of an earlier draft. I am also grateful to Ali M. Parhizgari for his able and energetic research and programming assistance I alone am, of course, responsible for the final presentation.

attention on two particular aspects of these programs: what we might call the "output side," that is, the kinds of service these programs were intended to deliver, and the extent to which the groups receiving the services—the poor themselves—were permitted to participate in the design and administration of the program.

Curiously, little attention has been paid to what I shall call the "input side"—namely, direct, salaried employment in these activities This myopia is surely true for the Model Cities Program One can find numerous stories in the popular press about the local conflicts surrounding the election of neighborhood boards, but one is hard pressed indeed to find any discussion of the manpower aspects of Model Cities Yet, if I am correct in my estimate that—by the middle of 1971—there were perhaps 25,000 salaried employees in the Model Cities Program across the country, then Model Cities as a source of jobs is now larger than New Careers, Operation Mainstream, or the Job Corps. Even if Model Cities fails to provide either high-quality services to the ghetto community or a vehicle for increased political participation by the urban poor (and in saying this, I do not mean to imply that such has been the case), it would still be an important part of our overall antipoverty effort simply because of the employment and wages it has generated

There are three kinds of positions in the Model Cities Program (1) direct City Demonstration Agency (CDA) jobs for professionals, paraprofessionals, clerical/office workers, technicians, laborers, and operatives (e g , truck drivers) on the central CDA staff itself, (2) similar employment in the municipal or private nonprofit "delegate agencies" (DAs) which are operating projects supported in whole or in part by HUD–Model Cities funds, and (3) a share of the jobs generated by new construction associated with the program (this category has produced too few jobs to permit analysis) If all of the estimated 25,000 jobs in fact went to the urban poor, then—at roughly $8,500 a job—ghetto income would be increased by $213 million a year, a not-insignificant sum

In fact, however, all of the jobs are *not* going to ghetto (i e , "model-neighborhood") residents On April 30, 1969, after nearly 2 years of "start-up," only 44 percent of the salaried employees in the 77 "first-round" Model Cities projects were from the designated target neighborhoods The "resident participation rate" (*RPR*) was only 22 percent for professionals (a not-unexpected result for a planning period), 68 percent for subprofessionals, 49 percent for clerical workers, and 80 percent for other categories Moreover, nonresidents (of all races) received substantially higher salaries than residents in the same occupations and with similar demographic characteristics; these salary differentials ranged from $800 to $1,300 per year (Harrison 1973).[1] A much more comprehensive survey of 75 projects in the spring of 1971—the data base for the present paper—indicates the absence of any improvement in the situation, again, fewer than half of all the jobs in the program were held by ghetto residents [2] Professional *RPR*s were 23 percent for whites and 30 percent for minorities For subprofessionals, technicians, clerical workers, and laborers, the respective *RPR*s were 63 percent and 74 percent, 57 percent and 51 percent, 46 percent and 52 percent, and 56 and 64 percent

---

[1] This study of the Model Cities planning period was conducted for the National Civil Service League with questionnaires administered by HUD and processed by the Bureau of Applied Social Research at Columbia University.

These employment practices violate both the intent of Congress and administrative regulations of the Department of Housing and Urban Development Title I of the original Demonstration Cities and Metropolitan Development Act of 1966 (PL 89-754), in authorizing HUD to undertake the Model Cities Program, stipulated that "residents of the model neighborhoods should also be given maximum opportunity for jobs in the projects and activities of the program " Section 103(a) of the act requires that each CDA provide "maximum opportunities for employing residents of the area in all phases of the program and enlarging opportunities for work and training " In 1970, the HUD assistant secretary for community development issued "CDA Letter Eleven," requiring participating cities to give preferential treatment to model-neighborhood residents as a necessary condition for retaining HUD–Model Cities funds (to date, no city has been cut off)

The present paper has as its purpose the explanation of intercity variation in total and occupation-specific *RPR*s, and the estimation of the magnitude of wage discrimination against model-neighborhood residents Probably the most important institutional factor informing the design of the econometric models that follow has been the lack of independence of Model Cities projects from their local city governments Unlike (for example) the Community Action Program of OEO "All funds for a Model Cities Program are directed to a local City Demonstration Agency (DCA) through the elected city council This gives local city councils final veto power over the program by ruling out any direct funding relationships between the community and HUD It is an inescapable conclusion that the official intention was to make city hall responsible for the participation of citizens and to prevent the development of autonomous, government-financed citizen organizations, as had occurred in the poverty program" (International City Managers' Association 1972, pp 12–13) This is why the specification of explanatory variables measuring the structure of city government receives the greatest priority in the following pages [3]

---

[2] The second Model Cities Survey was conducted during the winter and spring of 1971 by the National Civil Service League, on a contract from HUD The *RPR*s were generally higher in DAs than in CDAs—quite the opposite of what one would expect considering that CDAs are supposed to directly represent the interests of the model neighborhood For all occupations, the respective averages were 56 and 43 percent

[3] In the 1969 study, civil service and "potential black power" variables proved to be especially significant determinants of resident employment in Model Cities

As with the Community Action Program (CAP), the employment aspect of Model Cities has a variety of manifest functions· to give participants (model-neighborhood residents) direct economic benefits and skills, to "encourage" city agencies to become more responsive to the needs of the poor, and to challenge "credentialism" as a barrier to the public employment of the disadvantaged [4] Moreover, the actual delivery of services by city agencies is also involved. There is increasing evidence that, in nonwhite urban communities in particular, "no delivery of services may be possible unless there is participation" (Altshuler 1970, p. 221) [5]

Again, like CAP, there are also latent functions associated with the employment activities of Model Cities These include the creation of "demonstration effects" within the community, the development of an institution around which community organization is at least theoretically possible, the creation of a "fifth column" for the ghetto inside city hall, and the focusing of pressure on a system (public personnel) which is already under attack from other groups such as the still predominantly white unions

Evaluations of CAP, while equivocal to pessimistic about the impact of the program on local government behaviour, have been fairly optimistic in their assessment of the direct employment impact Dale Rogers Marshall writes "Poor people get jobs or positions in groups which lead to jobs They also exhibit increases in political education, information about how the system works, and in organizational skills and feelings of political efficacy Thus participation serves as a training ground for indigenous leaders" (Marshall 1971, p 473)

The Correlates of Resident Participation

The political science and sociology literatures include a number of quantitative studies of intercity variation in the performance of poverty programs Marshall, for example, writes that "differences in the implementation of participation are associated with differences among the communities in which [Community Action Agencies] are located Size of the city, size of the minority population, and political factors seem to make a difference" (Marshall 1971, p 458) A Wisconsin Poverty Institute study by Michael Aiken and Robert Alford regressed the local per capita poverty grant in 1966 on five types of independent variables size and age of city as proxies for the degree of bureaucratization, the present manifestations of bureaucratization (such as the presence of a civil service personnel system), the degree of poverty (as a measure of "need"), ethnic/cultural heterogeneity (which the federal government might take to be a predictor of conflict over the grant), and the power of the chief executive

---

[4] "Credentialism" is defined as the *excessive* (i e , job task–unrelated) use of educational credentials by employers as employee-selection devices, perhaps to facilitate the practice of "statistical discrimination" (cf Berg 1970, Ruchlin 1971, Harrison 1972, esp chap 1, on statistical discrimination, see Piore 1971)

[5] Thus, in its technical assistance efforts for HUD, the National Civil Service League has been stressing the use of "selective certification" to achieve hiring preference for model-neighborhood residents, on the grounds that cities cannot meet their legal and constitutional obligations to provide an equal distribution of services to all communities unless they use community residents to deliver services to their own neighborhoods

For Aiken and Alford (1968), the object of study was the relationship between the "ability to mobilize" (as measured by the five clusters of variables) and one particular measure of the payoff to mobilization—acquisition of a federal grant In the present paper, the problem to be explored is how the "ability to mobilize" affects the extent to which the urban poor benefit ("participate") from those grants in one particular way—through direct, paid employment

Specification

We define the resident participation rate as

$$RPR = \left(\frac{\text{number of employees who are model-neighborhood residents}}{\text{total number of employees}} \right) \times 100$$

Six different multiple regressions will be fitted, one for each of six *RPR*s corresponding to (1) all employees, (2) professionals only, (3) subprofessionals, (4) technicians and operatives, (5) clerical workers, and (6) laborers The mean values of these six variables were 48 percent, 26 percent, 66 percent, 55 percent, 48 percent, and 60 percent, respectively

The first class of explanatory variables to be examined measures actual Model Cities inputs One independent variable will be cumulative funding as of December 31, 1970 (X_1) The mean funding level of the 64 projects for which sufficient data were available to permit a complete analysis was $4 18 million It is difficult to predict how this would vary with *RPR* An alternative formulation is per capita funding (X_2), that is, total funding divided by the total number of project employees The mean value of (X_2) was $30,040

The second class of independent variables measures the structure of the municipal government The variable (X_3), the age of the city (as an indicator of its degree of bureaucratization), is measured by the last two digits of the year (after 1900) when the city reached a population size of 50,000, a threshhold selected after examining the size distribution of the 64 cities in the sample [6] The mean value of (X_3) is 27 (corresponding to the year 1927). Following Aiken and Alford, we should expect that the degree of bureaucratization increases with time, so that (as measured here) age and degree of bureaucratization are inversely related. It is my guess that city bureaucratization constitutes an obstacle to ghetto residents, bureaucratization and *RPR*s are probably also inversely related Therefore, the maintained hypothesis is $\partial RPR/\partial X_3 > 0$

---

[6] Cities which attained this scale before 1900 were arbitrarily assigned values of "00" Cities which had not yet attained a scale of 50,000 by 1970 were arbitrarily assigned values of "80" Data availability dictated this procedure

A more static variable is (X_4), the 1970 population level in the city (mean = 314,562) To the extent that larger cities contain more workers with whom ghetto dwellers must compete (and if they also contain larger and/or more strongly entrenched bureaucracies), size and *RPR* are likely to be negatively related

There are three different ways to measure the size of the bureaucracy directly by the 1968 level of municipal employment (X_5), by the percentage growth in municipal employment between 1962 and 1968

$$(X_6) = \frac{\begin{pmatrix}\text{municipal employment} & - & \text{municipal employment} \\ \text{in 1968} & & \text{in 1962}\end{pmatrix}}{(\text{municipal employment in 1962})} \times 100,$$

and by the proportion of the working residents of the city who are employed by the municipal government, estimated (somewhat crudely) by

$$(X_7) = \left(\frac{\text{1962 municipal employment}}{\begin{matrix}\text{1960 employment of all workers} \\ \text{living in the city}\end{matrix}}\right) \times 100$$

(Again, data availability forced this rather peculiar specification However, since public employment has been growing much more rapidly than private employment in cities [Harrison 1971], the fact that $[X_7]$ is undoubtedly biased upward for 1960 makes it a better indicator for the more recent past) The effect of increases in (X_5)–(X_7) on *RPR* is ambiguous Larger (stronger?) city bureaucracies are more capable of effectively blocking ghetto demands On the other hand, the faster a city's payroll grows, the easier it ought to be to accommodate the poor—especially in jobs subsidized by the federal government [7] The mean values of (X_5)–(X_7) are 4,797, 12 8 percent, and 7 7 percent, respectively

Still another way of measuring the extent of bureaucratization involves indicators of the degree to which local personnel practices are institutionalized Variables now susceptible of measurement include the presence or absence of a civil service/tenure system (X_8), whether or not written examinations are normally administered in hiring for unskilled vacancies (X_9), whether or not arrest records are used as selection instruments (X_{10}), and whether or not the jurisdiction has a full-time personnel director (X_{11}) These are all given dummy variable specifications; their mean values across the sample (interpreted as proportions) are 0 69, 0.29, 0.74, and 0 91, respectively Since nearly half of the Model Cities projects already operate within their local civil service systems—and all are being urged by HUD to do so—we should expect (X_9) and (X_{10}) to be inversely related to *RPR* (given what we know about the relatively poor cognitive skills and illegal activities of—as well as police attitudes toward—many ghetto residents, cf Harrison 1972, chaps 1, 5) I would expect the same to be true for the other two variables as well

---

[7] Piven (1972) believes that the rapid growth of city payrolls is to a great extent not a cause but a result of (or response to) white demands for sufficient employment to maintain "social distance" from blacks, who have unquestionably been increasing their absolute participation in municipal employment through the poverty programs Under this hypothesis, the level and/or rate of growth of city employment would still be positively correlated with *RPR*, although the direction of causation would be reversed.

In a study of the Model Cities and Community Action Programs in Chicago and Detroit, Allan Rosenbaum (1970) showed that, especially because of "the ambiguity of the federal intention," both programs have been redirected "towards the ends of the mayors of the two cities As a result the programs have, in most instances in these two cities, been simply mirror-image reflections of the political-institutional arrangements and environment in which they operated " The financial plight of big-city mayors has been well documented Given this constant pressure from influential constituents to locate external resources for local uses, and assuming that pressure on the mayors from ghetto organizations and individuals is relatively less pronounced, then the presence in a city of a strong mayor (or city manager) might well correlate negatively with *RPR*, as Model Cities funds are "leaked" to other (nonghetto) constituents or are substituted for local revenues (which were already being allocated for the employment of nonghetto residents) Three indicators of the relative political power of mayors can be quantified Let

$$(X_{12}) = \begin{array}{l} 1 \text{ if city has a mayor/council form of government,} \\ 0 \text{ otherwise,} \end{array}$$

$$(X_{13}) = \begin{array}{l} 1 \text{ if mayor has veto power over his/her council,} \\ 0 \text{ otherwise,} \end{array}$$

$$(X_{14}) = \begin{array}{l} 1 \text{ if mayor is in his/her second or greater term,} \\ 0 \text{ otherwise} \end{array}$$

The mean proportions for these three indicators of mayoral power are 0.71, 0 23, and 0.41, respectively All three are expected to be inversely related to *RPR* (the veto-power variable will actually be specified as $X_{13} \times X_{12}$)

The third set of explanatory variables measures actual or potential "black power," that is, the relative importance of minorities (not only blacks) in the city A wide variety of indicators can be specified, including:

$$(X_{15}) = \begin{array}{l}\text{model neighborhood population in 1965}\\ (\text{mean} = 32{,}873),\end{array}$$

$$(X_{16}) = \begin{array}{l}\text{1 if Citizens' Advisory Board was elected by the}\\ \quad\text{community,}\\ \text{0 otherwise (mean} = 0\ 72),\end{array}$$

$$(X_{17}) = \begin{array}{l}\text{1 if the local CDA director is a member of a minority}\\ \quad\text{group,}\\ \text{0 otherwise (mean} = 0.48),\end{array}$$

$$(X_{18}) = \text{1970 nonwhite city population (mean} = 73{,}052),$$

$$(X_{19}) = \begin{array}{l}\text{1970 nonwhite share of total city population}\\ (\text{mean} = 20 \text{ percent}),\end{array}$$

$$(X_{20}) = \frac{\left(\begin{array}{c}\text{1970 nonwhite}\\ \text{population}\end{array} - \begin{array}{c}\text{1960 nonwhite}\\ \text{population}\end{array}\right)}{(\text{1960 nonwhite population})} \times 100$$
$$(\text{mean} = 53 \text{ percent}),$$

$$(X_{21}) = \frac{(\text{1970 nonwhite share} - \text{1960 nonwhite share})}{(\text{1960 nonwhite share})} \times 100$$
$$(\text{mean} = 35 \text{ percent}),$$

$$(X_{22}) = \begin{array}{l}\text{1 if city had a major civil disorder in 1967,}\\ \text{0 otherwise (mean} = 0\ 42)\end{array}$$

It seems reasonable to expect that a large model neighborhood will be in a position to mobilize greater political support for a program such as Model Cities than will a smaller target population Similarly, the act of holding periodic elections for the local Citizens' Advisory Boards required by HUD may both manifest and reinforce community concern for a "piece of the action" (such concern may not produce results, of course but a negative correlation between $[X_{16}]$ and *RPR* is most unlikely)[8] The CDA director is appointed by the mayor or city manager, th selection of a minority director probably manifests a mayoral commitmen to the equity objectives of the program Thus, it is expected tha $\partial RPR/\partial X_{17} > 0$ Variables (X_{18})–(X_{21}) measure the level and growt of relative potential minority voting power in the city, all are expected t have a positive effect on *RPR* Finally, (X_{22}) measures an important negative sanction which ghetto dwellers are able (albeit often at great personal cost to themselves) to impose on city hall. The Kerner Commission's index of major civil disorders in 1967 provides the source for quantification of this "riot" variable Many social scientists attribute the expansion of antipoverty programs after 1967 to the political tension generated by the riots; if so, then we should expect that $\partial RPR/\partial X_{22} > 0$

---

[8] Boards must be formed by law, but they need not be elected. As indicated in the te over a quarter of the boards are in fact appointed. For analyses of the composition a activities of these Boards, see International City Managers' Association (1971) a Harrison (1973).

> One final variable deserves a word of explanation there is some evidence from earlier studies that unskilled urban workers in general and minorities in particular tend to fare relatively better in high wage than in low wage cities A high wage city generally contains an elaborate industrial base which consumes considerable quantities of labor of all skill types The large pool of labor made possible by this industrial base provides an "agglomeration economy" to new employers—in this case, the CDA In other words, we hypothesize that model neighborhood workers will get more Model Cities jobs in high wage cities because, in such cities, they are more likely to have had work experience [Harrison 1973, p 52]

In the 1969 study, this variable,

$$(X_{23}) = \left(\frac{\text{manufacturing wage bill in 1963}}{\text{manufacturing man-hours in 1963}}\right)$$
$$(\text{mean} = \$2\ 49),$$

was indeed highly significant for the laborer category. As will be seen, it was among the most important variables in the present study as well

Analysis

Initial inspection of the results of fitting the six regression models,

$$RPR_{in} = b_0 + \sum_{j=1}^{23} b_{jn}X_{jn} + \varepsilon,$$
$$i = 1, \quad , 6 \text{ occupation groups}$$
$$n = 1, \quad , 64 \text{ cities}$$

indicated the presence of substantial multicollinearity With a linear specification and multiple R's of 30– 80, first-order (pairwise) correlation coefficients of more than .30 crudely indicate intercorrelation of sufficient magnitude to generate potentially serious multicollinearity. Experimentation with different combinations of variables led to the decision to delete variables (X_2), (X_4), (X_5), (X_{12}), (X_{15}), (X_{18}), and (X_{19}) (the selection of a variable to represent a collinear set was determined by ease of policy interpretation). Four theoretically important variables—city age, the presence of a civil service system, growth of the black population, and growth of the black population share—were retained despite some collinearity by running separate pairs of regressions for each of the six *RPR*s, including (X_3, X_{20}) and (X_8, X_{21}), respectively.

$$\begin{aligned} RPR_{in} = {} & b_0 + b_1X_1 + b_3X_3 + \sum_{j=6}^{7} b_jX_j \\ & + \sum_{j=9}^{11} b_jX_j + \sum_{j=13}^{14} b_jX_j + \sum_{j=16}^{17} b_jX_j \\ & + b_{20}X_{20} + \sum_{j=22}^{23} b_jX_j + \varepsilon, \\ & \qquad i = 1, \quad , 6 \\ & \qquad n = 1, \quad , 64 \end{aligned} \tag{1}$$

$$\begin{aligned} RPR_{in} = {} & b_0 + b_1X_1 + b_6X_6 + \sum_{j=7}^{11} b_jX_j \\ & + \sum_{j=13}^{14} b_jX_j + \sum_{j=16}^{17} b_jX_j \\ & + \sum_{j=21}^{23} b_jX_j + \varepsilon \\ & \qquad i = 1, \ldots, 6 \\ & \qquad n = 1, \quad , 64 \end{aligned} \tag{2}$$

The results are displayed in table 1 It is immediately apparent that employees in the 64 Model Cities projects for which complete data were available fall into two very distinct groups, in terms of the relationships under study The proportions of professional and subprofessional hires drawn from the ghetto are much less strongly related to the political-economic structure of the city than are the corresponding *RPR*s for technicians/operatives, clerical workers, and laborers, as indicated by the large differences in regression coefficients This dual stratification of the Model Cities labor force was confirmed by pooling the professional-subprofessional observations and the technician-clerical-labor observations, fitting regressions to each data set, and conducting Chow tests on the *F*-statistics for these regressions and one fit to the entire sample. The null hypothesis that the two structures come from the same population was rejected at the 99 percent level of confidence For this reason, the all-occupation results have been deleted from table 1.

These differences are worth exploring further. It is true that the participation of ghetto residents in all five occupation groups is positively promoted by the rate of growth and relative importance of municipal employment (variables X_6 and X_7).[9] Moreover, subprofessionals at least share with the nonskilled labor categories the circumstance of being relatively disadvantaged by the presence in a city of a civil service/tenure system (X_8) and a strong mayor, as measured by whether or not the latter has a veto over his or her council (X_{13}). On the other hand, professional ghetto residents gain in relative employment from civil service (X_8) and written exam (X_9) propensities in the municipality; the other groups lose. This is especially true for laborers, for whom a city's decision to use written exams for unskilled jobs is associated with a decrease of fully 5 percent in their resident participation rate. Moreover, while a number of the "black power" variables—including the race of the CDA director (X_{17}), the growth of the black population level (X_{20}) and share (X_{21}), and whether or not the city experienced a riot in 1967 (X_{22})—positively affected the *RPR*s of the second group of ghetto workers, they were either unrelated to the *RPR*s of professionals and subprofessionals, or else—in the case of (X_{21})—had a negative effect, implying that the increasing relative importance of blacks in a city's population *endangers* the relative opportunities of the middle class in the ghetto.

Within the second group of occupations, the funding variable (X_1) is positively associated with *RPR* for technicians/operatives and laborers (two categories whose members are nearly all males). Apparently, it is the projects which are the best funded that most intensively employ model-neighborhood unskilled or semiskilled ("working-class") men. These are also the two categories of sample workers who are the most likely to have acquired police arrest records. For them, a jurisdiction's propensity to use such records as screening devices reduces their chances of employment in the program, as expected. The relationship between (X_{10}) and the technician/operative and laborer *RPR*s is negative and highly significant.

For the three nonskilled occupations, the government-structure variables behave quite consistently (although age of city [X_3] was never statistically significant, perhaps because of its restricted range). Both "mayoral power" variables (X_{13}, X_{14}) contribute to reduced resident participation, as hypothesized. Whether or not the city had a full-time personnel officer was of no consequence.

As indicated before, all three nonskilled occupational groups benefit from increases in actual or latent "black power." The "riot" variable (X_{22}) displays larger *t*-statistics than any other variable in the model,

---

[9] Still, the professional and subprofessional coefficients tend to be rather larger than those for the other occupations, the respective ranges are 0.962–2.371 and 0.593–1.113 for (X_6) and 1.372–1.451 and 0.056–0.088 for (X_7). As the municipal demand for labor grows, ghetto professionals and subprofessionals perhaps face relatively less competition from "outsiders" than do workers in the other occupations because of the relative shortages of skilled (or specialized) labor residing in (or available to) the city as a whole. These shortages do not in general obtain for the other occupational categories.

TABLE 1
Coefficients Estimated from Resident Participation Regressions

Independent Variable	*RPR* Professionals (1)	*RPR* Professionals (2)	*RPR* Subprofessionals (1)	*RPR* Subprofessionals (2)	*RPR* Technicians/ Operatives (1)	*RPR* Technicians/ Operatives (2)	*RPR* Clerical (1)	*RPR* Clerical (2)	*RPR* Laborers (1)	*RPR* Laborers (2)
	Model Cities Variables									
1 Funding level ($00000)	0.097 (0.092)	0.099 (0.093)	0.062 (0.073)	0.064 (0.073)	0.592* (0.240)	0.511* (0.213)	0.513 (0.321)	0.621 (0.393)	0.422* (0.175)	0.468* (0.155)
	Government-Structure Variables									
3 Age (1900 = 00, 1970 = 70)	0.162 (0.169)		0.188 (0.173)		0.068 (0.065)		0.072 (0.081)		0.063 (0.072)	
6 Percentage change in municipal employment (1962–68)	0.997* (0.230)	0.962* (0.161)	2.363* (0.091)	2.371* (0.092)	1.110* (0.437)	1.113* (0.438)	0.602* (0.241)	0.600* (0.240)	0.593* (0.242)	0.612* (0.259)
7 Municipal employment (1962) as percentage of all employment (1960)	1.428* (0.414)	1.451* (0.416)	1.372* (0.502)	1.382* (0.509)	0.065 (0.035)	0.056* (0.025)	0.088* (0.037)	0.070* (0.024)	0.085* (0.032)	0.079* (0.030)
8. Civil Service/tenure (0–1)		1.802* (0.885)		−1.750* (0.621)		−0.944 (0.661)		−2.421* (0.741)		−2.893* (0.640)
9. Written exams used for unskilled jobs (0–1)	2.201* (0.275)	2.003* (0.251)	−1.062 (0.951)	−1.001 (0.940)	−2.218* (0.940)	−2.212* (0.939)	−2.742* (0.861)	−2.535* (0.882)	−5.100* (1.749)	−5.006* (1.658)
10 Police records used for unskilled jobs (0–1)	2.345 (2.116)	2.012 (2.002)	−1.343 (1.260)	−1.212 (0.982)	−3.750* (1.020)	−3.640* (0.971)	−0.301 (0.210)	−0.272 (0.185)	−6.444* (1.139)	−6.017* (1.012)
11 City has full-time personnel officer (0–1)	−0.922 (0.617)	−1.031 (0.692)	−0.630 (0.655)	−0.671 (0.662)	−0.821 (0.850)	−0.910 (0.874)	−0.644 (0.510)	−0.710 (0.531)	−0.775 (0.812)	−0.823 (0.841)
13 Mayor has veto over council (0–1).	−1.118 (0.912)	−1.243 (1.021)	−0.821* (0.402)	−0.862 (0.483)	−1.388* (0.615)	−1.375* (0.610)	−1.740* (0.602)	−1.817* (0.714)	−1.412* (0.510)	−1.435* (0.542)
14 Mayor is in second or greater term (0–1)	−1.421 (0.883)	−1.062 (0.688)	−0.945 (0.722)	−0.982 (0.741)	−3.450* (1.009)	−3.512* (1.017)	−2.843* (1.039)	−2.615* (1.002)	−1.989* (0.872)	−1.972* (0.716)

Table 1 (*Continued*)

	Black Power Variables									
16. Citizens' board is elected (0–1)	−0.828	−0.930	1.212	1.220	1.421	1.408	0.916	0.878	1.235	1.343
	(0.730)	(0.725)	(0.880)	(0.892)	(1.020)	(1.020)	(0.762)	(0.744)	(0.981)	(1.019)
17. CDA director is minority (0–1)	−0.642	−0.663	0.912	0.861	3.112*	3.222*	2.877*	3.012*	4.044*	4.263*
	(0.712)	(0.692)	(0.910)	(0.870)	(0.360)	(1.034)	(0.943)	(1.020)	(0.812)	(1.004)
20. Percentage change in black population (1960–70)	−1.435		−0.947		0.160*		0.510*		0.326*	
	(0.794)		(0.512)		(0.058)		(0.200)		(0.140)	
21. Percentage change in black share of city population (1960–70)		−0.120*		−0.083*		0.285*		0.470*		0.433*
		(0.035)		(0.040)		(0.115)		(1.021)		(0.110)
22. Major civil disorder in 1967 (0–1)	−0.921	−0.946	0.787	0.810	5.606*	5.640*	6.014*	6.210*	6.002*	6.245*
	(0.658)	(0.663)	(0.519)	(0.601)	(0.712)	(0.713)	(0.919)	(1.003)	(1.230)	(1.251)
	Climate for Unskilled Labor									
23. Average hourly manufacturing wage in city (1963)	1.213	1.102	1.030	1.410	3.742*	3.612*	4.310*	4.137*	4.742*	4.617*
	(1.118)	(1.093)	(0.912)	(0.988)	(0.941)	(0.882)	(1.205)	(1.200)	(1.656)	(1.640)
$\bar{R}^2$ (14,49 df)	0.231	0.202	0.110	0.090	0.657	0.604	0.703	0.682	0.715	0.700

SOURCES.—FOR VARIABLES
1—1970 HUD Statistical Yearbook, table 2
2—1970 Census, Series PC(1)–A, table 7
6—1962 Census of Governments, vol. 3, Compendium of Public Employment, table 26, International City Managers' Association, Urban Data Service, May 1969, vol. 1, no. 5, tables 22, 23
7—1962 Census of Governments, 1967 County and City Data Book, table 4
8—International City Managers' Association, *1967 Municipal Yearbook*, tables 6, 7
9—National Civil Service League, unpublished survey data
10—National Civil Service League, unpublished survey data
11—National Civil Service League, unpublished survey data
13—1968 Municipal Yearbook, table 2
14—1967, 1969, and 1971 Municipal Yearbooks, tables 14, 13, 1/4, respectively
16—Model Cities Administration, HUD, private communication
17—HUD-NCSL Model Cities Survey
20—1960 Census, vol. 1, table 13, 1970 Census, Series PC(1)-B, tables 23, 27, 32
21—1960 Census, vol. 1, table 13, 1970 Census, Series PC(1)-B, tables 23, 27, 32
22—National Advisory Commission on Civil Disorders, U.S. Riot Commission Report (1968), p. 158
23—1967 County and City Data Book, table 4

NOTE.—Standard errors in parentheses
* Significantly different from zero at the .05 level

and its quantitative effect on the *RPR* of nonskilled ghetto workers ranges between 5.6 and 6.3 percent. It would appear highly likely that those city halls which experienced major civil disorders in 1967 have been more responsive to the employment demands of unskilled ghetto workers.

Finally, the manufacturing wage variable (X_{23}) again displays a positive and significant relationship to resident participation for nonskilled workers. This is consistent with the hypothesis that high-wage *cities tend to contain a large "secondary labor market" which employs* large numbers of unskilled and clerical workers (Gordon 1972, Harrison 1972, chap. 5). To the extent that (even unstable) work experience enhances a ghetto dweller's chances for employment in the Model Cities Program—whether because of the recruiting practices of the CDA, job search practices of the resident workers, or both—we should expect to find a positive relationship between *RPR* and a proxy for the degree of work experience of the unskilled ghetto labor force. The average wage rate may be such a proxy.

Wage Discrimination against Model-Neighborhood Residents

The survey of 75 Model Cities projects indicated that both part-time and full-time nonmodel-neighborhood residents earn higher incomes than residents of the target neighborhoods. For whites, the ratio of median resident/nonresident annual earnings ranged (across all occupations or job categories) from 76 to 97. The range for minority employees (blacks, Puerto Ricans, chicanos, American Indians, orientals, and "others") was 57–106. The first-order correlations between residence (= 1 if *MN*, 0 otherwise) and earnings were −0.623 for whites and −0.647 for minorities, both coefficients were significantly different from zero at the 99 percent level of confidence.

Multiple regression analysis was employed in order to explore several possible explanations for this apparent wage discrimination against model-neighborhood residents. Differences in earnings might be attributable to the relatively lower incidence of possession by ghetto residents of educational credentials required for "admission" to municipal jobs. A complementary explanation—proposed most recently by Barbara Bergmann (1971)—suggests that neighborhood residents are "crowded" disproportionately into the lowest-paying jobs. This is likely to be especially prevalent for female employees.

Two linear microregressions were estimated for the 10,176 white and two for the 4,131 minority employees. The results are displayed in table 2.[10] Clearly, model-neighborhood residents earn systematically lower

---

[10] Separate models for examining the relationship between earnings and education on the one hand and earnings and occupation on the other were necessitated by the high correlation between education and occupation.

TABLE 2
EMPLOYEE WAGE REGRESSIONS

INDEPENDENT VARIABLE	DEPENDENT VARIABLE ANNUAL EARNINGS ($)			
	Whites		Minorities	
	(1)	(2)	(1)	(2)
X_1 = Model-neighborhood resident (0–1)	−1237.70* (20.07)	−1179.36* (22.76)	−1445.65* (67.34)	−1311.26* (69.62)
X_2 = Male (0–1)	524.26* (20.64)	516.05* (23.13)	929.64* (69.60)	860.79* (52.98)
X_3 = Age (years)	232.65* (18.78)	228.47* (19.70)	189.23* (23.46)	163.84* (23.23)
X_4 = Age squared	−6.14* (0.25)	−7.68* (0.26)	−4.23* (0.30)	−3.19* (0.29)
X_5 = Full-time worker (usually ≥35 hours/week) (0–1)	1049.91* (21.25)	1045.06* (22.92)	622.05* (95.24)	638.03* (104.15)
X_6 = Civil service/tenure coverage (0–1)	−17.93 (25.13)	−9.30 (26.54)	−102.14 (80.31)	−148.33 (79.54)
X_7 = Completed exactly 12 years of school (0–1)	80.80 (54.73)		−34.43 (270.02)	
X_8 = Completed exactly 13–15 years of school (0–1)	−64.61 (46.38)		1030.85* (271.58)	
X_9 = Completed 16 or more years of school (0–1)	2551.21* (237.38)		1826.55* (258.35)	
X_{10} = Professional (0–1)		2287.68* (41.03)		1441.13* (128.75)
X_{11} = Subprofessional (0–1)		820.07* (39.22)		1021.12* (122.33)
X_{12} = Technician/operative (0–1)		−37.50 (67.50)		−404.94 (209.99)
X_{13} = Clerical worker (0–1)		114.56* (15.23)		248.35 (142.79)
X_{14} = Employed directly by the CDA (0–1)	2716.07* (58.20)	2813.60* (61.31)	3266.46* (156.23)	3609.09* (135.81)
X_{15} = Employed by a delegate agency producing education services (0–1)	83.98* (35.20)	77.90* (35.03)	579.28* (151.05)	550.83* (142.79)
X_{16} = Employed by some other type of delegate agency[a] (0–1)	21.71 (32.71)	22.34 (34.72)	573.41* (136.82)	532.17* (131.43)
Degrees of freedom	12, 10163	13, 10162	12, 4118	13, 4117
$\bar{R}^2$	0.846	0.828	0.518	0.530

SOURCE.—HUD-NCSL Model Cities Survey

NOTE.—Standard errors are in parentheses

[a] To avoid singularity in the $[X'X]$ matrix, a category of miscellaneous delegate agencies employing 12% of all Model Cities workers was removed from the set of agencies reflected in variable X_{16}

* Significantly different from zero at the .05 level

TABLE 3

EXPECTED ANNUAL EARNINGS FOR TWO REPRESENTATIVE MODEL CITIES WORKER TYPES (IN DOLLARS)

	Nonresident Accountant	Resident Teacher's Aide
White:		
Eq (1)	$13,651	$6,706
Eq (2)	13,263	7,364
Minority		
Eq (1)	12,194	4,805
Eq (2)	11,379	5,729
Difference		
Eq (1)	1,457	1,901
Eq (2)	1,884	1,635

salaries than nonresidents, even after the effects of differential education, occupation, and "industry" (CDA, education-oriented delegate agency, or other delegate agency) have been removed Among whites, the expected differential in earnings attributable to residence is $1,179–$1,238 per year, for minorities, the differential is even larger $1,311–$1,446 per year Neither occupational "crowding" nor differences in years of schooling provide sufficient causal explanations, in fact, experiments with interaction terms such as (X_1X_7), (X_1X_8), and (X_1X_9) showed the marginal effect of education on earnings to be rather insensitive to residence[11]

This wage discrimination against model-neighborhood residents is compounded by additional wage discrimination against minorities[12] Consider, for example, two representative worker types (*a*) a 30-year-old, nonresident, college-educated, male professional employed full-time by the CDA (say, as an accountant), and (*b*) a 30-year-old resident female subprofessional with only a high school diploma, working full-time for a local school board (say, as a teacher's aide) Substitution of the appropriate values into the fitted regressions yields the expected earnings estimates displayed in table 3. The differences between annual earnings by race are remarkably similar for two such disparate worker types.

---

[11] The raw data permit the identification of 23 specific occupations (e g, "subprofessional-law enforcement") and 23 employing agencies (e g, "fire department"). Analysis of regressions specified at this level of detail show that earnings vary nearly as much within as among occupations and agencies—suggesting that the crowding hypothesis as a theory of wage determination may not apply to the public sector For regressions including (*a*) neither occupations nor agencies, (*b*) occupations only, (*c*) agencies only, and (*d*) occupations and agencies, the respective coefficients of variation were .17, 15, 15, and 12

[12] The mean white wage was $9,023; the mean minority wage was $6,912 The sample of neighborhood, i e, ghetto, residents is not biased toward minorities Forty-nine percent of the 10,176 whites lived in the target areas, while 51 percent of the 4,131 minority employees were model-neighborhood residents

about $1,500–$1,900 for the nonghetto accountant and $1,600–$1,900 for the ghetto teacher's aide. It is, of course, theoretically possible that these differences are in fact attributable to some factor(s) not specified in the model, rather than to race per se. However, with between 52 and 85 percent of the wage variation explained by the specified models—a very high degree of "explanatory power" for cross-section regression analysis—it is likely that at least part of these estimated differences are indeed attributable to discrimination

Summary and Conclusions

The Model Cities Program of the U S. Department of Housing and Urban Development, although it is principally concerned with the delivery of services to the urban ghetto, has in fact become one of the largest manpower programs operating in urban labor markets In 1971, HUD subsidized some 25,000 budgeted positions in 903 municipal and nonprofit agencies, at an average salary of nearly $8,500 Two-thirds of the participating agencies provided on-the-job and other forms of training to Model Cities employees

Both the original legislation and subsequent administrative rulings by HUD require each participating city to give preference in hiring to residents of the target areas—the urban ghetto Yet, by the spring of 1971, fewer than half of all the jobs in the program were in fact held by ghetto residents Moreover, nonresidents were being paid significantly higher salaries than residents for performing essentially the same tasks

This paper has attempted to identify the determinants of intercity variation in "resident participation rates"—the share of jobs held by ghetto residents—and the causes of the resident/nonresident differences in earnings

In the resident-participation regressions, particular attention was addressed to variables reflecting the political structure of the cities in the program—including degree of bureaucratization (e g., civil service variables) and mayoral political strength (as measured by, e g , number of terms and veto powers)—and actual or latent "black power" (e g , whether or not the city experienced a major civil disorder in 1967) In general, cities with highly structured bureaucracies, strong mayors, and formal civil service systems and procedures allocate a larger share of Model Cities jobs to "outsiders" (in violation of the federal guidelines) than is the case for other cities On the other hand, real or potential black political power increases ghetto participation in the Model Cities Program, as does the presence of strong private-sector demand for unskilled labor (the latter effect is probably related to the propensities of Model Cities agencies to use previous work experience as an employee-selection criterion)

Analysis of 14,307 individual employee records shows that neither occupational "crowding" nor lack of educational credentials sufficiently explains the salary differences between ghetto and nonghetto Model Cities workers, after controlling for sex, age, employment status, civil service coverage, and type of employing municipal or nonprofit agency Moreover, we were able to identify what appears to be an independent discriminatory effect by race

I must conclude that, in terms of the legislative mandate for giving priority in Model Cities employment to the residents of the target areas themselves, these findings are disheartening The conclusions of my earlier study is no less valid here· "Once again, the urban poor have not received as large a share as might have been possible of the benefits associated with a government program ostensibly directed to improving their welfare This 'leakage' of public anti-poverty resources into the hands of individuals at least some of whom are of doubtful standing as proper recipients has characterized other federal programs It is unfortunate that Model Cities appears to have begun its career in a similar way" (Harrison 1973, p 54)

These results also bear on another—and much larger—issue of public policy the question of decentralization of manpower programs The research on which this paper reports suggests (although it of course does not prove) that local governments with the political power to do so—those with strong bureaucracies and powerful chief executives—can and will distort the equity objectives of federal programs Findings such as these therefore have a useful role to play in future discussions about the reorganization of manpower and antipoverty efforts

References

Aiken, Michael, and Alford, Robert R "Community Structure and Mobilization The Case of the War on Poverty " Discussion Paper no 29–68, Univ Wisconsin, Inst Res Poverty, October 1968

Altshuler, Alan *Community Control· The Black Demand for Participation in Large American Cities.* New York Pegasus, 1970.

Berg, Ivar. *Education and Jobs· The Great Training Robbery* New York Praeger, 1970

Bergmann, Barbara R. "The Effect on White Incomes of Discrimination in Employment " *JPE* 79 (March/April 1971)· 294–313.

Gordon, David M *Theories of Poverty and Underemployment.* Lexington, Mass.· Heath-Lexington, 1972

Harrison, Bennett. *Public Employment and Urban Poverty.* Washington· Urban Inst. 1971.

———. *Education, Training, and the Urban Ghetto.* Baltimore: Johns Hopkins, 1972

———. "The Participation of Ghetto Residents in the Model Cities Program." *J. American Inst. Planners* 39 (January 1973): 43–55.

International City Managers' Association. "Citizens and Model Cities: Major Issues and Relationships." In *1971 Municipal Year Book*. Chicago: Internat. City Managers' Assoc., 1972.

Marshall, Dale Rogers. "Public Participation and the Politics of Poverty." In *Race, Change, and Society*, edited by Peter Orleans and William Russell Ellis, Jr. Los Angeles Sage, 1971.

Piore, Michael J. "The Dual Labor Market." In *Problems in Political Economy. An Urban Perspective*, edited by David M. Gordon Lexington, Mass · Heath, 1971.

Piven, Frances Fox "Cutting Up the City Pie" *New Republic*, February 5, 1972, pp. 17–22

Rosenbaum, Allan "Participation, Programs, and Politics The Federal Impact on the Metropolis" Paper delivered to the 1970 meetings of the American Political Science Association, Los Angeles

Ruchlin, Hirsch S. "Education as a Labor Market Variable" *Indus Relations* 10 (October 1971) 287–300

Optimum City Size: The External Diseconomy Question

J. V. Henderson

Queen's University, Canada

The paper discusses whether market-achieved city size is greater or less than optimum city size when externalities such as pollution are present. Optimally taxing or controlling pollution may or may not lead to a decline in production of polluting goods Moreover, taxation reorders production and consumption patterns in the city so that the minimum net effect on welfare of city inhabitants is the decline in pollution Since pollution taxation increases the welfare of city inhabitants, city size rises with immigration from the rest of the economy

A popularly held idea is that cities, in particular big cities, are too large and that optimum size cities would be smaller [1] Nonoptimality arises because of unpriced production or consumption externalities such as noise, air, or water pollution and congestion, which if properly controlled would lead to a reduction in city sizes This paper disputes this conclusion for numerous reasons For illustrative purposes, the paper concentrates on an analysis of air and water pollution

The first task is to formulate the problem in a simple but appropriate context Firms produce an export good in a city Due to a production input such as sulphur-bearing coal or petroleum, the production of the export good results in air- or water-borne emissions such as sulphur dioxide, carbon by-products, and phenols [2] This pollution may affect production efficiency of firms or households, or it may just enter utility

---

The theme and the general approach to a few of the problems discussed in the paper were inspired by an unpublished paper by George Tolley The author is further indebted to Oded Israeli and anonymous referees for helpful written comments on the paper and is indebted to members of the Urban Economics Workshop at the University of Chicago for verbal comments on the June 1972 presentation of this paper

[1] For example, see Sandquist (1970) See Mills and de Ferranti (1970) for a discussion of the subject This second article reaches no firm conclusions on this subject.

[2] See Kneese (1971) for a description of the multiple forms and effects of pollution

functions of people living in the city. Concerning this, one might hypothesize that since many work places are air-conditioned and purified primarily to stabilize seasonal fluctuations in temperature, air pollution in general does not enter production functions or affect efficiency significantly. Similarly, water pollution little affects the efficiency of water use in industrial processes. However, pollution certainly affects consumption activity, either directly entering the utility function or affecting efficiency of household production. That is, it is part of the amenity income from living in a city.

What happens if these externalities are priced to equate social and private marginal costs (assuming they can be priced)?[3] The usual analysis states that, if production of a good results in external diseconomies, optimally taxing this good will result in a decline in its production and an increase in production of other goods. The inference is then made that if production of the polluting good falls, the size of cities producing that good falls. In contrast, my analysis leads to the following conclusion. Optimally pricing pollution will lead to a reordering of production priorities in cities and the economy away from goods responsible for pollution. Within the city, the adverse effect of reduced production of an export good will be more than offset by the increase in output of other goods and the decline in pollution. Welfare of city residents will rise, and immigration to the city from the rest of the economy will then occur.

Before proceeding to the main analysis, the concept of a city and of a system of cities in an economy must be examined. In discussing optimum city size, it is very important to understand what a city is and does and how it relates to the total economy.

Cities produce a bundle of goods. As characterized by Mills (1967), in even the most simple model, cities produce an export good or bundle of goods, housing, and a commuting good. Agglomeration in cities occurs because of scale economies in producing, say, the export good. Workers work in the city center producing the export good and commute from their home site to the city center and back each day. As city size increases, per person commuting costs rise due to increasing spatial distances that workers must commute and rising congestion on the roads. As expressed in Mills (1967) and in Henderson (1972), these rising commuting costs with agglomeration may offset the benefits of agglomeration or scale economy exploitation in export good production. That is, there is an efficient city size for production and consumption.

Extending this analysis, Henderson (1972) then shows if there is an efficient city size where net benefits of agglomeration are maximized,[4] there may be multiple cities producing the same bundle of goods.[5] Further, as has been asserted by numerous people (e g , Lösch 1954 and Beckmann 1968), Henderson also shows there are different types of cities in the economy. Each type of city specializes in the production of different good or bundle of export goods.[6]

---

[3] See Baumol and Oates (1971) and Bohm (1971) for a description of pricing schemes and other methods of pollution control.

[4] Henderson does not state that the market achieves this efficient city size. His analysis and any reasons for nonoptimality are independent of the analysis in this paper or external diseconomies.

These propositions indicate there is a system of cities in an economy with different types of cities producing different bundles of export goods. Within each type of city, there may be one or multiple cities depending on economy or world demand for the export goods of that city type, their costs of transportation in trade (the Löschian [1954] limitation on city size), and their production technology, in particular, the degree of scale economies (the extent of city agglomeration economies) In discussing the effect upon city size of taxing pollution in cities producing polluting goods, later in the paper we will take into account the relationship between these cities and other types of cities in the economy

Externalities in Consumption

In this section, we formally specify the nature of pollution and examine what happens to city size if pollution is optimally priced Our analysis is carried out in terms of a representative city in the economy of the type that produces a polluting good Implications of our analysis for the country as a whole rather than just the representative city are examined later in the paper

Our representative city produces X, an export good, and Y, a non-traded good or bundle of goods such as housing and retail services The production functions for the X and Y industry are, respectively,

$$X = X(L_X, N), \tag{1}$$

$$Y = Y(L_Y) \tag{2}$$

where L_X and L_Y are labor inputs and N is the input of an imported natural resource These functions are assumed to be locally homogeneous of degree one As explained above, to have urban agglomeration, either

---

[5] This result can be derived in a Löschian (1954) world where efficient city size is limited by the increasing transport costs of supplying larger and larger rural areas as city size increases

[6] Henderson's (1972) reasoning is as follows If the production of two export goods or bundles or export goods involves no benefits from a common location such as utilizing a common specialized labor force or locationally fixed intermediate input, then it may be beneficial to locate the production of these goods in different cities Locating their production together increases the labor force that must be housed and that adds to average commuting costs in the city while not increasing efficiency in the production of the traded goods Separating the production of the two goods into different cities lowers the increase in per person commuting costs relative to scale economy exploitation per unit of traded-good output, as traded-good production and city size rise. Separating the two industries of course involves further transportation expenditures due to intercity trade, and a complete analysis would have to account for these costs

X or Y would globally have to have increasing returns to scale Local linear homogeneity could imply (dis)economies of scale are exhausted, at least at current city size [7] Note that urban scale economies are also needed to explain why people would cluster together in a city where there is pollution which they could avoid by spatially moving apart We will comment on these scale economy problems below

The fact that firms in the X industry use an input N, such as coal or petroleum, results in pollution such as sulphur dioxide or carbon emissions Firms using N all contribute flows of emissions that result in a "stock" of pollution This stock then enters people's utility functions or firm production functions as an unpriced public good (bad) The pollution production function is $p = P(N_1, N_2, \ldots, N_n)$, where N_i are N inputs of the ith firm in the X industry and where $\partial P/\partial N_i \gtreqless P/N_j$ Note that firms' marginal contributions to the stock of pollution, $\partial P/\partial N_i$, may differ depending on relative spatial locations of firms and prevailing winds or on individual firm production functions In this section we assume all $\partial P/\partial N_i$'s are equal or pollution is a function of the industry use of N

$$P = g(N) \tag{3}$$

The production function of P is not homogeneous of degree one At low levels of X output, additions of N add little to pollution At higher levels the natural absorption of pollution by the atmosphere or water due to chemical breakdown or spatial distribution by wind or water currents is slowed, and additions of N contribute significantly to pollution This formulation of pollution being a "stock" created by emission flows from individual firms is due to Tolley (1969)

Pollution is currently unpriced and enters people's utility functions as a negative amenity Since pollution is caused by use of N, this means N is improperly priced, that is, units of N should be taxed according to their indirect effect through pollution on people's utility What happens to output and city size when units of N and pollution are taxed? Recall that N is an imported resource The social marginal cost of N lies above its import price By taxing N and reducing its use, one feels intuitively that people in the city should be made better off since we are reducing the use of an input for which social cost exceeds expenditures or perceived private benefits (marginal product) Just how city inhabitants are made better off is examined below rigorously both to emphasize the interactions involved and to prove the point In particular, we are curious as to how production and consumption patterns change, how wages change, and what the effect of returning pollution tax proceeds to city inhabitants is.

---

[7] Why is it, for example, that major cities such as New York, Boston, Chicago, Los Angeles, etc keep growing at a steady rate comparable to the national population growth rate? Since these cities expand neither more rapidly nor less slowly than an average increase, neither can their scale diseconomies be overwhelming nor can their scale economies be unfulfilled I am suggesting they have reached a stage where city output as a whole has constant returns to scale Just how that breaks down for different sectors is another question

In setting up the problem, we assume city inhabitants have equal utility levels with inhabitants in other cities or parts of the economy We assume initially following the pollution tax that city inhabitants remain in the city It is the usage of N and output and consumption patterns that adjust first to the pollution tax Following this adjustment, if the welfare of city inhabitants rises, then migration to this city from the rest of the economy should occur over time Essentially we are solving a problem that is a "partial equilibrium" problem with respect to the rest of the economy To solve rigorously for an economy general equilibrium solution, one would want to have a total model of an economy solving for all city sizes and factor usages where pollution is initially unpriced and then priced in those cities where there is pollution. Algebraically this is not feasible,[8] but I believe the arguments in this paper will indicate that the directional changes in city size in my model would persist in a full solution

To examine the adjustment of the city to the pollution tax, we set up a simple general equilibrium model of a city The production functions for the city are given in equations (1), (2), and (3) On the consumption side, it is assumed all consumers have identical tastes and incomes City population, L, is equated to city labor force, $L_X + L_Y$, for simplicity Per person utility is defined by

$$U_i = U_i(X^c/L, Y/L, Z^c/L, P) \qquad i = 1 \quad m \tag{4}$$

For X and Z, superscripts, c, refer to goods consumed in the city as opposed to those produced in the city The city exports $X - X^c$ and imports Z^c and the input N The balance of trade equation for the city is

$$p_X(X - X^c) = p_Z Z^c + p_N N, \tag{5}$$

or city income (which equals $p_L L$ before pollution taxes) equals

$$I = p_X X + p_Y Y - p_N N, \tag{6}$$

where p's refer to prices The p_X and p_Z are determined in national or international markets and are fixed for the city (see discussion below)

When N is taxed for pollution, the tax proceeds are redistributed to laborers in the city Then per laborer income, I/L, is

$$I/L = p_L - \left(\frac{p_t}{L}\right) P = p_X\left(\frac{X^c}{L}\right) + p_Y\left(\frac{Y}{L}\right) + p_z\left(\frac{Z^c}{L}\right) \qquad p_t < 0, \tag{7}$$

---

[8] Henderson (1972) solves for a general equilibrium model of an economy with cities That solution is complicated enough without adding in externalities and taxation solutions to externalities The model simply becomes unsolvable

where p_t is the price per unit of pollution and $(p_t/L)P$ is the negative per person head tax. If we maximize utility in equation (4) subject to the budget constraint in equation (7), we can calculate the optimal pollution tax or tax on N, which we will call t. From the constrained maximization problem, the Samuelson public-good condition says

$$p_t/p_X = \frac{U_P L}{U_X}, \tag{8}$$

where $U_X = \partial U/\partial X$, $U_P = \partial U/\partial P < 0$. Equation (8) states that the price ratio equals the sum of the marginal utilities of the public good normalized by the marginal utility of a private good. To determine the relationship of the pollution price to the tax on N, we first examine the relationship between head taxes (pollution prices) and taxes on N and then use the production relationship for P in equation (3). If P and N were the original quantities of pollution and N, then

$$p_t(P - \Delta P) = -t(N - \Delta N), \qquad t > 0, \tag{9}$$

which says that proceeds from the post-tax employment of N equal subsidies (or negative head tax) paid to laborers, where Δ's refer to changes. Rearranging and using equations (3), (8), and (9), we get the tax

$$t = -(1 - \overset{*}{N})^{-1}(P/N - g_N\overset{*}{N})\left(p_X \frac{U_P}{U_X} L\right) > 0,^{9} \tag{10}$$

where an asterisk refers to rates of change and $g_N = dg(N)/dN = dP/dN$.

We can now turn to examining the effect of the tax, t, upon equilibrium in the city. To determine the change in equilibrium, we employ equations of change from Jones's (1965) paper on general equilibrium models. The production side of the model is given by Jones's equations of full employment and equations indicating price equals marginal (average) cost or there is perfect competition

$$\begin{bmatrix} a_{XL} & a_{YL} \\ a_{XN} & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} L \\ N \end{bmatrix}, \tag{11}$$

$$\begin{bmatrix} a_{XL} & a_{XN} \\ a_{YL} & 0 \end{bmatrix} \begin{bmatrix} p_L \\ p_N \end{bmatrix} = \begin{bmatrix} p_X \\ p_Y \end{bmatrix}, \tag{12}$$

where a_{ij} from activity analysis refers to the amount of the jth input used in a unit of the ith output.

From equation (12), we want first to obtain price changes that may occur on the production side of the model. We assume that the price of X is fixed in national or international markets and that the import price of N

---

[9] $t > 0$, since $P/N > g_N\overset{*}{N} = \frac{dP}{dN}\frac{dN}{N}$, $\overset{*}{N} < 1$, and $U_P < 0$.

is constant. The purchase price of N rises by the amount of the tax, t If the purchase price of N rises and p_X is fixed, then p_L, the price of labor, must fall to keep the cost of producing X from rising If p_L falls, then p_Y, the price of the nontraded good, must fall in perfect competition since labor is the only factor of production in Y These results may be obtained formally by differentiating equation (12) and solving to yield[10]

$$\overset{*}{p}_L = \overset{*}{p}_Y = -\frac{\theta_{XN}}{\theta_{XL}}\overset{*}{p}_N < 0, \tag{13}$$

where asterisks refer to rates of change (e g , $\overset{*}{p}_L = dp_L/p_L$) and θ_{ij} is jth factor's share in the revenue of the ith good and $\overset{*}{p}_N = t/p_N$

The price changes in equation (13) induce changes in factor inputs, N, L_X or $a_{XL}L$, and L_Y or $a_{YL}L$ in equation (11) Using an elasticity of substitution defined as $\sigma_X = (\overset{*}{a}_{XL} - \overset{*}{a}_{XN})/(\overset{*}{p}_N - \overset{*}{p}_L)$ and the relationships in equation (13), we differentiate equation (11) and simplify to get

$$\begin{bmatrix} \lambda_{XL} & \lambda_{YL} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \overset{*}{X} \\ \overset{*}{Y} \end{bmatrix} = \begin{bmatrix} 0 \\ \overset{*}{N} \end{bmatrix} + \sigma_X \overset{*}{p}_N \begin{bmatrix} -\lambda_{XL}\,\theta_{XN}/\theta_{XL} \\ 1 \end{bmatrix}, \tag{14}$$

where λ_{ij} is the fraction of the jth factor used in production of the ith good (note $\lambda_{XN} = 1$ and $\lambda_{YN} = 0$) Initially $\overset{*}{L} = 0$ by assumption

Changes in output are determined by changes in factor inputs suggested in equation (14) and by demand conditions Demands for X^c, Y, and Z^c are determined by income and prices Since p_X and p_Z (the price of the import good) are fixed in national markets, to solve for changes in equilibrium output (not consumption) we only need one demand equation, $Y = Y(I, p_X, p_Y, p_Z)$ We differentiate this equation, substituting in $\overset{*}{p}_X = \overset{*}{p}_Y = 0$ and substituting for $\overset{*}{I}$ from equation (6) where it is assumed that all pollution tax proceeds are returned to city inhabitants This yields[11]

$$\overset{*}{Y}(1 - k_Y\eta_{Y,I}) = \frac{p_X X}{I}\overset{*}{X} - \frac{p_N N}{I}\eta_{YI}\overset{*}{N} + \eta^S_{Y,p_Y}\,\overset{*}{p}_Y, \tag{15}$$

---

[10] Differentiating equation (12) yields

$$\begin{bmatrix} \theta_{XL} & \theta_{XN} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \overset{*}{p}_L \\ \overset{*}{p}_N \end{bmatrix} = \begin{bmatrix} 0 \\ \overset{*}{p}_Y \end{bmatrix} + \begin{bmatrix} \theta_{XL}\overset{*}{a}_{XL} + \theta_{XN}\overset{*}{a}_{XN} \\ 0 \end{bmatrix},$$

where, since N is only used in X production, $\overset{*}{a}_{YL} = 0$, $\theta_{YL} = 1$, $\theta_{YN} = 0$ $\overset{*}{p}_X = 0$ by assumption Jones (1965) further reasons that $\theta_{XL}\overset{*}{a}_{XL} + \theta_{XN}\overset{*}{a}_{XN} = 0$ by principles of cost minimization. Given that, solving the equation yields equation (13)

[11] To get equation (15) we use the relationships $\overset{*}{Y} = \eta_{YI}\overset{*}{I} + \eta_{Y,p_Y}\overset{*}{p}_Y$, $\eta^S_{Y,p_Y} = \eta_{Y,p_Y} + k_Y\eta_{Y,I}$, $k_Y = p_Y Y/p_L L = p_L L_Y/p_L L = \lambda_{YL}$, and $\overset{*}{I} = -(p_N N/I)\overset{*}{N} + (p_X X/I)\overset{*}{X} + (p_Y Y/I)(\overset{*}{Y} + \overset{*}{p}_Y)$.

where η_{YI} and η^S_{Y,p_Y} are the income and compensated price elasticities of demand for Y and k_Y is the fraction of consumer income spent on Y.

Combining equations (15) and (14) and using $\overset{*}{p}_Y = -(\theta_{XL}/\theta_{XL})\overset{*}{p}_N$ yields the equations determining output changes.

$$\begin{bmatrix} \lambda_{XL} & \lambda_{YL} & 0 \\ 1 & 0 & -1 \\ \frac{-p_X X}{I}\,\eta_{YI} & (1 - k_Y\eta_{Y,I}) & \eta_{YI}\,\frac{p_N N}{I} \end{bmatrix} \begin{bmatrix} \overset{*}{X} \\ \overset{*}{Y} \\ \overset{*}{N} \end{bmatrix}$$

$$= \overset{*}{p}_N \begin{bmatrix} -\lambda_{XL}\,\theta_{XN}/\theta_{XL} \quad \sigma_X \\ \sigma_X \\ -\eta^S_{Y,p_Y}\,\theta_{XN}/\theta_{XL} \end{bmatrix} \tag{16}$$

Solving using Cramer's rule we get

$$\overset{*}{X} = \overset{*}{p}_N \left(\frac{\theta_{XN}\lambda_{YL}}{\theta_{XL}\lambda_{XL}}\,\eta^S_{Y,p_Y} - \frac{\theta_{XN}}{\theta_{XL}}\,\sigma_X \right) < 0, \quad \text{since } \overset{*}{p}_N > 0,\ \eta^S_{Y,p_Y} < 0, \tag{17}$$

$$\overset{*}{Y} = \overset{*}{p}_N \left(-\frac{\theta_{XN}}{\theta_{XL}}\,\eta^S_{Y,p_Y} \right) > 0, \tag{18}$$

$$\overset{*}{N} = \overset{*}{p}_N \left(-\frac{\sigma_X}{\theta_{XN}} + \frac{\lambda_{YL}\theta_{XN}}{\lambda_{XL}\theta_{XL}}\,\eta^S_{Y,p_Y} \right) < 0 \tag{19}$$

Since pollution, P, equals $g(N)$ from equation (3), for $\varepsilon_N = (dP/dN)(N/P)$

$$\overset{*}{P} = \varepsilon_N \overset{*}{N} < 0 \tag{20}$$

Note all our results are in terms of the $\overset{*}{p}_N$ variable (t/p_N) whose positive directional sign we know

The decline in X production and N usage is expected since the production price of N rises The increase in Y production (equals consumption) is expected since p_Y declines Note, however, these changes are only a function of the elasticity of substitution in production and the compensated elasticity of demand of Y The former indicates the degree to which N declines in X production with an increase in p_N and hence indicates the degree to which pollution may be reduced by substitution in production The second indicates only substitution effects are present in consumption of market goods We have taxed N which, given a fixed p_X, means p_L and hence p_Y must decline. But the tax proceeds from N offset the income effect of the price and wage changes to laborers, leaving only pure substitution effects in consumption of market goods.

To solve for changes in utility levels of city residents we first must solve for consumption changes of X^c and Z^c We differentiate the balance-of-trade equation (5) $[p_X(X - X^c) = p_Z Z^c + p_N N]$, and substitute in for $\overset{*}{X}$, $\overset{*}{N}$, and then $\overset{*}{Z}{}^c$ by differentiating the demand equation for $Z^c = Z(I, p_X, p_Y, p_Z)$. We solve first for $\overset{*}{X}{}^c$ and then for $\overset{*}{Z}{}^c$.

$$\overset{*}{X}{}^c = \overset{*}{p}_N\left(-\frac{\theta_{XN}}{\theta_{XL}}\,\eta^S_{X,p_Y}\right) < 0, \qquad \text{iff } \eta^S_{X,p_Y} > 0, \tag{21}$$

$$\overset{*}{Z}{}^c = \overset{*}{p}_N\left(-\frac{\theta_{XN}}{\theta_{XL}}\,\eta^S_{Z,p_Y}\right) < 0, \qquad \text{iff } \eta^S_{Z,p_Y} > 0 \tag{22}$$

Not surprisingly, X^c and Z^c decline with the fall in the price of Y providing all goods are substitutes in consumption

To solve for the change in per person utility we differentiate equation (4) to get

$$dU = X^c \cdot \partial U/\partial X^c\left(\frac{dX^c}{X^c}\right) + Y \cdot \partial U/\partial Y \cdot \left(\frac{dY}{Y}\right)$$
$$+ Z^c \quad \partial U/\partial Z^c\left(\frac{dZ^c}{Z^c}\right) + P \quad \partial U/\partial P\left(\frac{dP}{P}\right)$$

We then substitute in the first-order conditions for a maximum—for example, $\partial U/\partial X^c = \lambda p_X$ where λ is the marginal utility of income—and substitute in for $\overset{*}{X}{}^c$, $\overset{*}{Y}$, $\overset{*}{Z}{}^c$, and $\overset{*}{P}$ from equations (21), (18), (22), and (20) This yields

$$\frac{dU}{\lambda} = \overset{*}{p}_N\left[-p_L\frac{\theta_{XN}}{\theta_{XL}}\left(k_Y\eta^S_{Y,p_Y} + k_X\eta^S_{X,p_Y} + k_Z\eta^S_{Z,p_Y}\right)\right.$$
$$\left. + P\left(\frac{\partial U}{\partial P}\frac{1}{\lambda}\right)\varepsilon_N\left(-\frac{\sigma_X}{\theta_{XL}} + \frac{\lambda_{YL}}{\lambda_{XL}}\frac{\theta_{XN}}{\theta_{XL}}\,\eta^S_{Y,p_Y}\right)\right]$$

But $\sum_i k_i\eta^S_{i,p_j} = 0$

$$dU/\lambda = \overset{*}{p}_N\left(P\frac{\partial U}{\partial P}\,1/\lambda\right)\varepsilon_N\left(-\frac{\sigma_X}{\theta_{XL}} + \frac{\lambda_{YL}}{\lambda_{XL}}\frac{\theta_{XN}}{\theta_{XL}}\,\eta^S_{Y,p_Y}\right) > 0,$$
$$\text{since } \frac{\partial U}{\partial P},\ \eta^S_{Y,p_Y} < 0 \tag{23}$$

Therefore, the effects on utility of reduced X^c and Z^c are offset by the increased Y That is, in terms of market goods, the effects on utility net out to zero Since P is the only other argument in the utility function, as indicated in equation (23), the dollar value of utility increases by the dollar value of the reduction in pollution

As mentioned at the beginning, intuitively this makes sense The basic city resource is labor, which produces goods for its own consumption and to trade for other goods We tax the use of an input the city imports. The effect of the tax is to reorder production priorities in the city away from X, where marginal social cost exceeds price, to Y Since the tax proceeds are returned to labor, only substitution effects in market consumption are present, and we showed that they net out in terms of affecting utility. The net result is a Pareto-optimal reduction in the imported input and the accompanying positive benefits from the optimal reduction in pollution. Given that utility levels in the city rise, city size should increase over time as workers from other parts of the economy migrate to the city.

Three comments on the above analysis are in order First, allowing p_X to vary would not affect our results As X exports from the city fall, p_X could rise in national markets relative to its initial level This will affect the change in consumption and production patterns in the city, but not the basic effect on utility of a Pareto-optimal reduction in pollution Second, we could solve our model with import controls on N as the method of pollution control Again, production and consumption patterns will differ from the tax solution, but the net effect on utility will only be a fall in pollution

Our third comment is more serious in nature In setting up the problem, we assumed the X and Y production functions were locally linear homogeneous If we were to consider situations where $f_i(L_i, N_i)$ and $F(L_Y)$ were not locally homogeneous of degree one, our result could be weakened. If the production of X were to have industry increasing returns to scale and Y to have decreasing returns to scale, then the reduction of X and increase in Y would have adverse effects on labor productivity and the production potential of the city For example, if one postulates an export base theory for the existence of cities the fall in X output could have quite adverse effects on city size On the other hand if Y or the service sector of the city were responsible for the urban scale economies, the result of rising utility levels would be reinforced as production priorities were reordered to the non-polluting service sector Moreover an interesting situation arises in the case where there are production externalities or where pollution affects firm or household production efficiency In that case production of both goods, X and Y, may rise following the imposition of taxes We now turn to this situation

Externalities in Production

As before, our representative city produces an export good and a non-traded good or bundle of goods such as housing, retail services, and household produced goods (leisure, meals, child rearing, etc). Firms in the X industry utilize an input that results in pollution which affects production efficiency in the X and Y industries. Specifically, a firm's production function in the X industry is

$$x_i = g_i(P) f_i(L_i, N_i), \qquad i = 1, \ldots, n \tag{24}$$

$$\partial f_i/\partial N_i > 0, \quad \partial f_i/\partial L > 0; \quad (dg_i/dP)(\partial P/\partial N_j) < 0, \qquad i, j = 1 \ldots n,$$

where x_i is the firm's output; L_i and N_i are inputs of labor and the imported resource N; and P is pollution. The f_i is the (locally) linear homogeneous production function of the firm.[12] The $g_i(P)$ is a production function shifter[13] that affects firms' efficiency in a Hicks neutral fashion with respect to factor inputs; for example, pollution corrodes machinery as well as irritating workers Our specification is essentially one of industry scale diseconomies external to the firm.

The industry production function for Y is

$$Y = G(P)F(L_Y) \tag{25}$$

$$\partial F/\partial L_y < 0, \quad dG/dP \cdot \partial P/\partial N_i < 0, \qquad i = 1 \quad n$$

The $G(P)$ caused by usage of N in X production is also a production function shifter The pollution production function is as before

$$P = (N_1, \quad , N_n) \tag{26}$$

$$\partial P/\partial N_i \gtreqless \partial P/\partial N_j, \qquad i \neq j$$

To examine the economic effects of pollution on production and output in the city, we analyze optimum versus equilibrium factor pricing in the city. It is assumed readers are familiar with the two constrained optimization problems of firms maximizing profits versus an economy maximizing a social welfare function that determine equilibrium versus optimum pricing of factors We turn directly to the results In competitive equilibrium, factors are paid the value of their perceived marginal products That is,

$$p_L = p_X g_i(P)\ \partial f_i/\partial L_i, \qquad i = 1, 2 \quad n, \tag{27}$$

$$p_N = p_X g_i(P)\ \partial f_i/\partial N_i + p_X\ dg_i/dP \cdot \partial P/\partial N_i\ f_i, \qquad i = 1 \quad n, \tag{28}$$

where f_i is an abbreviation for $f_i(L_i, N_i)$ From equation (27) we can see that the marginal product of N_i to the firm is the sum of the normal first term plus the effect of using N_i upon pollution that then affects (negatively) the firm's own production Of course the firm cannot distinguish between the two effects and simply sees their sum as the marginal product of N.[14] For Y, from equation (25)

---

[12] See above discussion on the local linear homogeneity assumption

[13] Our production specification is chosen for analytic convenience In the alternative specification, P is an unpriced input in the f_i and F functions in equations (15) and (16) Then in addition to the externality distortion caused by the X industry emitting pollutants, we have an unpriced input distortion in the X and Y industries, since P would then be an unpriced input in the f_i and F production functions This would mean that P might earn a negative rent which, in perfect competition in output and factor markets, would have to be "assigned" to the earning of one or both of the other factors creating further distortions such as analyzed by Worcester (1969) and Meade (1952). There has been some debate on how to deal with the rent to an unpriced factor (see, for example, Mishan 1965), and we avoid this debate by using the production function specification where the only distortion is the externality itself

$$p_L = p_Y G(P)\, dF/dL_Y. \tag{29}$$

Equations (27) and (24) also describe the value of social marginal products for labor. However, the value of social marginal products for N are different from equation (28). The N_i affects production efficiency of firms in the X and Y industries by contributing to pollution which in turn acts as a production function shifter in these industries. For firms j and k the social marginal products of N are

firm j.

$$\begin{aligned} SMP_j &= p_X(g_j(P)\ \partial f_j/\partial N_j + dg_j/dP \cdot \partial P/\partial N_j\ f_j) \\ &\quad + \left[p_X\left(\sum_{\substack{i=1\\ i\neq j}}^{n} dg_i/dP \cdot \partial P/\partial N_j\ f_i\right) + p_Y\ dG/dP\ \ \partial P/\partial N_j\ \ F\right], \end{aligned} \tag{30}$$

firm k

$$\begin{aligned} SMP_k &= p_X(g_k(P)\ \partial f_k/\partial N_k + dg_k/dP\ \ \partial P/\partial N_k \cdot f_k) \\ &\quad + \left[p_X\left(\sum_{\substack{i=1\\ i\neq k}}^{n} dg_i/dP\ \ \partial P/\partial N\ \ f_i\right) + p_Y\ dG/dP\ \ \partial P/\partial N_k\ \ F\right], \end{aligned} \tag{31}$$

where f_i and F are abbreviations for $f_i(L_i, N_i)$ and $F(L_Y)$. The first term in the square brackets indicates the negative effect in the form of a production function shift of employing additional N_i on other firms in the X industry, and the second term represents the negative effect in the Y industry. In each case, per unit use of N should be taxed by the amount in the square brackets to equate both factor price with social marginal product and social and private marginal costs. The standard result following the imposition of the tax is that, with the increase in p_N, the use of N, the use of L in X (for p_X fixed), and output of X fall.

What does this indicate with respect to city size? To answer this question, I assume as above that the city labor force remains initially fixed and ask what happens to the welfare of city inhabitants following the imposition of the tax. If the welfare rises, the presumption is that people from other parts of the country will migrate to the city over time.

Although the imposition of the tax leads to a fall in X output, the fall in N usage means the city has to spend less of its income on importing N. In fact, since the value of social marginal product of N (eqs. [30] or [31]) is less than the perceived value of private marginal product of N which equals the price of N (eq. [28]), reductions in N mean that the reduction in city income spent on N (p_N multiplied by the change in N) should exceed the fall in the value of output lost by reductions in N This relationship should persist as long as the price of N exceeds the value of the social marginal product of N[15] Therefore income and welfare of city inhabitants will rise following taxation

---

[14] The adding-up problem is avoided by assuming that $p_X dg_i/dP\ \ \partial P/\partial N_i \cdot f_i$ is negligible. Then factor payments will exhaust firm revenue (see Chipman 1970).

As mentioned in the previous section, the proposition that, because pollution taxation corrects divergences between private and social costs, optimum city size will rise with pollution taxation may be simplistic For the case of externalities in consumption, we demonstrated that pollution taxation reorders production priorities and providing taxation proceeds are returned to city inhabitants, immigration should occur to the city The same type of analysis applied to externalities in production However, the same qualifications remain Pollution taxation may lead to a fall in the output of the good responsible for urban scale economies or, from an export base theory of city existence, the good responsible for city existence This would have adverse effects on city size if the assumption of local linear homogeneity is inappropriate

However, there are situations with externalities in production where no qualification is necessary In these situations when N is taxed and its usage falls, output of both X and Y may rise City inhabitants of course are better off, but also the export base and the service base of the city expand There are three pertinent situations

1 Pollution in a city may conceivably be so extensive that additional use of N may actually reduce total output Although $\partial f_i/\partial N_i > 0$, additions (reductions) of N may add (remove) so much to pollution and hence so affect production efficiency through large marginal reductions in the g_i and G terms in equations (24) and (25), that, in net, total output will fall (rise) In other words, in competitive equilibrium, due to the terms in the square brackets, equations (27) and (28) for the social marginal product of N may be negative If N is optimally priced or taxed, the importation and use of N will fall in the city For initial reductions in N total productivity will rise, because the true social marginal product of N is negative[16] In the social optimum, total output may rise and of course city income (value of output minus imports of N) will rise (Recall total output is maximized when the social marginal product of N is zero The social marginal product of N moves from an initial negative value to a final positive value, so it cannot be definitely concluded that total output rises) Whether X or Y both may rise or only Y rises depends, in addition to demand conditions, on whether the social marginal product

---

[15] Tolley (1969) essentially makes a similar point To quote Tolley "A manifestation of the gain from reducing the externality is likely to be that the production function shift lowering costs, is greater than the effect of the tax as a cost-raising item "

[16] See Gordon (1954) for a similar point involving fishing grounds. If the social marginal products are originally negative, then the original equilibrium would be inefficient (we are interior to the transformation curve) as well as nonoptimal (market pricing is incorrect).

of N in X production alone is originally negative—on whether the first term in the square brackets in equation (31) outweighs the private marginal product of N (the expression in parentheses in eq. [31]). Then both X and Y output may rise.

2 A simple example will illustrate our second situation There are two firms initially producing equal amounts of X, and both experience equal amounts of pollution However, only one firm creates pollution due to differences in production functions or spatial locations—for example, the "nonpolluting" firm may be downwind on the edge of the city If all production were shifted to the nonpolluting firm, for the same inputs, total output of X and Y would rise

In other words, firms may not contribute equally to pollution at the margin given either their production technology or relative spatial locations, that is, $\partial P/\partial N_j \neq \partial P/\partial N_k$ Alternatively, $\partial g_j/\partial P \neq \partial g_k/\partial P$, or pollution affects efficiency of firms differently for the same reasons If either is the case, equations (30) and (31) for social marginal product will not be equal in competitive equilibrium (Of course, private marginal products will be equal) Suppose the social marginal product of N in firm k is greater than in firm j If we held employment in Y and city labor force fixed, and moved some of N_j to firm k, output of X would rise for the same inputs

As shown by Schall (1971), all this implies that we are not on the Pareto-efficient transformation surface in competitive equilibrium where equations (28) but not (30) and (31) are equalized If we impose a Pareto-optimal tax on the use of N, we move from the private competitive equilibrium interior to the transformation curve to the social optimum on the transformation curve for the city By taxing N_i in the production of X, we (a) rearrange production in the X industry efficiently, shifting us to the transformation curve, and (b) eliminate the divergence between private and social marginal costs caused by the X industry polluting itself and the Y industry Usually only (b) is considered, which would indicate a fall in X production and a rise in Y production By including (a) where we rearrange N usage and X production between firms to those polluting less or being less affected by pollution, X output may rise despite the tax on N and its fall in total usage.

3 Our third situation is a more economically technical one and our discussion is brief As Plott (1966) points out, because pollution is caused by the use of an input, taxing the input N may still lead neither to a fall in X output nor in use of labor As shown in Bear (1965), this occurs if N is an "inferior" input or if N and L are "substitutes" in production.[17] Taxing and reducing the use of N and raising its price can lead to an increase in X output for p_X held constant.

---

[17] Inputs are substitutes in production if the addition of one input to production holding the second input constant reduces the marginal product of the second input.

The Representative City and the Economy

The result of taxing pollution in the representative city analyzed above was to make the existing workers there better off Given that, we argued that over time immigration would occur to the city Then city output of Y would rise even further and output of X would probably rise relative to its initial post-tax level If pollution is optimally taxed in all cities of the type that produce X, then sizes of all these cities should increase over time, even though economy output of X declines due to the rise in price of N These cities now have simply reordered production priorities from X to Y, the nontraded good such as retail services, housing, and household produced goods, optimum city size has risen because these cities are now more efficient production-consumption units

It seems even possible that X production in these cities may rise with immigration, with the appropriate changes in production techniques to less pollution-intensive methods of production [18] Of course, unless one of the three situations depicted in the production externalities section prevails, X output in the economy will still decline with the change in price of N For X production in some cities to rise and X output of the economy to fall, there must be a decline in the number of cities producing X That is, given initial multiple cities of the type that produce X, the optimum size of that type of city or that type of production-consumption unit has risen, and hence the numbers of that type of city will decline. Just what solution will prevail could only be determined in a model of a system of cities far more complex than any that currently exist

Conclusions

My conclusion is that optimally taxing an externality will make people in a city better off at the existing size of the city That will induce immigration to the city and increase city size The analysis in the paper is specific to pollution but can be extended

Immediate parallels for other types of externalities spring to mind, in particular for congestion Is the effect of optimally pricing congestion just to raise the price of peak-time travel and make it more expensive to commute? Or is the effect of pricing congestion to allocate resources efficiently, reordering travel times and modes Moreover, switching to a time-variable congestion price would most likely reduce off-peak travel times relative to current gasoline tolls and raise peak travel times Also, evaluation of optimal investment in roads and financing of roads would change with optimal pricing (see Mohring 1970)

Parallels to the situations in the production externality case in the paper also arise. Walters (1961) has shown that when traffic density becomes very high, traffic flow will decrease (this occurs at about 11 miles an hour). Optimal congestion taxes will lower traffic densities by reordering times when people travel and increase traffic flows or output! Furthermore, by taxing vehicles differentially according to their contribution to congestion, traffic composition may be changed at peak hours, increasing automobiles relative to trucks on intracity roads during rush hour That is, truckers may then choose to travel at nonpeak times and to travel the periphery of the city, rather than through the city

---

[18] That is, the N/L ratio will fall as more of the nonresource input is used for either pollution control or less polluting methods of utilizing the resource input.

References

Baumol, W J , and Oates, Wallace E "The Use of Standards and Prices for Environment Protection " *Swedish J Econ* , no 1 (Summer 1971), pp 42–54

Bear, D V T "Inferior Inputs and the Theory of the Firm " *J P E.* 73 (June 1965) 287–89

Beckmann, M *Location Theory* New York Random House, 1968

Bohm, Peter "An Approach to the Problem " *Swedish J Econ* , no 1 (Summer 1971), pp. 55–66

Chipman, J. S "External Economies of Scale and Competitive Equilibrium." *Q J E* 84, no 3 (August 1970) 347–85

Gordon, H S "The Economic Theory of a Common-Property Resource The Fishery " *J P E* 62, no 2 (April 1954) 124–42

Henderson, J V "The Types and Sizes of Cities A General Equilibrium Model." Ph D dissertation, Univ. Chicago, 1972

Jones, R W "The Structure of Simple General Equilibrium Models " *J P.E.* 73, no 6 (December 1965) 557–72.

Kneese, A. V "Background for the Economic Analysis of Environmental Pollution " *Swedish J Econ* , no 1 (Summer 1971), pp 1–24

Losch, A *The Economics of Location* New Haven, Conn Yale Univ Press, 1954

Meade, J E "External Economies and Diseconomies in a Competitive Situation " *Econ J* 62, no 245 (March 1952)· 54–67

Mills, E S "An Aggregative Model of Resource Allocation in a Metropolitan Area " *A E R* 57, no 2 (May 1967) 197–210

Mills, E S , and de Ferranti, D "Market Choices and Optimum City Size " *Papers and Proc American Econ Assoc* (1970)

Mishan, E J "Reflections on Recent Developments in the Concept of External Effects " *Canadian J Econ* 21 (February 1965) 3–34

Mohring, H "The Peak Load Problem with Increasing Returns and Pricing Constraints." *A E R.* 60, no 4 (September 1970) 693–705

Plott, C R "Externalities and Collective Taxes " *Economica* 33 (February 1966) 84–87.

Sandquist, J "Where Shall They Live?" *Public Interest,* no 18 (1970), pp. 88–100

Schall, L D. "Technological Externalities and Resource Allocation " *J P.E* 79, no. 5 (September-October 1971)· 983–1001.

Tolley, G. S. "The Welfare Economics of City Bigness " Urban Econ. Report no. 31, Univ. Chicago, 1969

Walters, A. "The Theory and Measurement of Private and Social Cost of Highway Construction " *Econometrica* 29 (October 1961)· 679–99.

Worcester, D A., Jr. "Pecuniary and Technological Externality, Factor Rents, and Social Costs." *A.E R* 59, no. 5 (December 1969). 873–85

Communications

A Conditional Logit Model of Occupational Choice

Michael J Boskin

Stanford University

I. Introduction

The theory of human capital[1] provides a convenient framework for analyzing the choice of occupation by individual workers The present paper is devoted to an investigation of the implications of this framework Section II briefly discusses the human capital approach to occupational choice Section III reviews the conditional logit statistical model elaborated in McFadden (1968) Section IV discusses the generation of the data Section V presents the empirical results of this study, that is, tests of hypotheses about the variables influencing occupational choice and estimates of the relative weights given to variables in selecting an occupation Section VI offers a brief conclusion

II. The Human Capital Approach to Analyzing Occupational Choice

The application of the theory of human capital to occupational choice is simple and straightforward [2] In choosing among occupations (defined broadly enough so workers in different occupations are not perfect

---

I would like to thank Daniel McFadden both for introducing me to conditional logit analysis and for the use of his computer program I also received valuable advice on this paper from Mel Reder, John Pencavel, Michael Hurd, Arthur Goldberger, Takeshi Amemiya, and the referee Thomas Moore provided computational assistance Finally, I am pleased to acknowledge the support of the Manpower Administration, U.S Department of Labor, through research grant 91-06-72-28 I am solely responsible for what remains

[1] See Schultz (1962) and Becker (1964)

[2] See, for instance, Fleisher (1970)

substitutes), a potential worker will weigh the benefits—potential earnings and nonpecuniary returns—and costs—of training, foregone earnings, etc. The worker will invest in changing occupations only if the returns are sufficiently large to make the particular change of occupation the most profitable use of his limited resources.

In this simplest formulation, we would always expect to find workers in those occupations with the highest discounted present value of future potential earnings.[3] However, given imperfect capital markets, resources for investing in oneself will not be equally accessible to all workers. The wealth position of an individual will partially determine his capability of making any particular profitable investment in himself

It is thus clear that decisions on occupation choice will be governed by the returns—primarily expected potential (full-time) earnings—and costs—primarily training and foregone potential earnings—relative to the wealth position of the individual worker in alternative occupations That is, the probability that a worker, i, will enter a particular occupation, j, will be a function of the relative present values of potential post-investment lifetime earnings, E, training costs and foregone earnings relative to wealth, T/W_i, and the present value of expected income foregone due to unemployment, U, in alternative occupations,[4] that is,

$$p_{ij} = f(E_{i1}, \quad , E_{ij}, E_{ij+1}, \quad , E_{in}, U_{i1}, \quad , U_{ij}, \quad , U_{in}, \\ T_{i1}/W_i, \quad , T_{ij}/W_i, \quad , T_{in}/W_i) \tag{1}$$

We turn now to a discussion of the estimation of such a relationship

III. A Conditional Logit Model of Occupational Choice[5]

If we assume that preferences are such that all occupations have a positive selection probability for each individual, that the odds that a particular occupation will be chosen over another are independent of the presence of other possible occupations, and that the determination of the odds of choosing alternative occupations is additively separable in its arguments representing the occupation effect and the individual worker effect, we may invoke certain theorems proved by McFadden (1968)

---

[3] Abstracting for the moment from any differential nonpecuniary costs and benefits among occupations More rigorously, the worker must choose simultaneously (1) the optimal number of occupations over his lifetime, (2) the optimal length of stay in each occupation, and (3) the optimal sequence of occupations That is, we have an integer dynamic programming problem Therefore, our worker will always be in an occupation which enables him to maximize present value considering all possible occupational shifts.

[4] This formulation may be motivated by an appeal to stochastic choice theory For example, following Block and Marshak (1960), we may adopt a model whereby the probability of each outcome is proportional to the utility derived from the choice

[5] The discussion that follows is based on McFadden (1968). This estimator has also been used to study choice of college by Radner and Miller (1970)

These theorems essentially allow us to write the selection probability in the form

$$p_{ij} = \left\{\sum_{k=1}^{n} \exp[b(k) - b(j)]\right\}^{-1} \tag{2}$$

The weight $b(k)$ assigned to an occupation is a function of the attributes of the occupation The individual observes a vector of variables, X_k, such as wage rates, unemployment rates, training costs, etc , for each occupation The $b(k)$ which determine the selection probabilities are given by

$$b(k) = B(X_k, \theta), \tag{3}$$

where θ is a vector of unknown parameters specifying the functional form of B. For each individual, define a variable $f_k = 1$ if occupation k is selected and $f_k = 0$ otherwise Then

$$\sum_{k=1}^{n} f_k = 1$$

To estimate θ, we observe that the likelihood of a given sample is then

$$L = \prod_{i=1}^{m} \prod_{k=1}^{n} p_{ik}^{f_k} \tag{4}$$

Then,

$$\log L = -\sum_{i=1}^{m} \sum_{k=1}^{n} f_k \log\left\{\sum_{k=1}^{n} \exp[B(X_k, \theta) - B(X_j, \theta)]\right\} \tag{5}$$

The method of maximum likelihood can be applied to (5) to obtain an estimator for θ with optimal asymptotic properties McFadden (1968) shows that the estimator for the case where X_i and θ are $K \times 1$ vectors and B is the linear function $B(X_i, \theta) = \theta' X$ is consistent and asymptotically normal This result is used to construct approximate large sample confidence bounds for the estimates

We thus have for our case of a linear function of occupation attributes that the ratio of the odds of choosing occupation j over occupation k is given by

$$\frac{P_{ij}}{P_{ik}} = \frac{e^{\theta' X_j}}{e^{\theta' X_k}}. \tag{6}$$

Taking logarithms yields

$$\log \frac{P_{ij}}{P_{iK}} = \theta'(X_j - X_i), \tag{7}$$

the log of the odds that occupation j will be chosen over occupation k is a linear function of the attributes of the occupation. Thus, we have the multinomial extension of traditional logit analysis (see Theil [1969] for another discussion of this statistical problem)

IV. The Data

We begin this discussion of the empirical results with an explanation of the generation of the data We estimate the present value of expected lifetime whole income (expected wages × hours available for work, assumed = 2,000) for individuals in our sample for various occupations they might enter These observations are taken from the 1967 Survey of Economic Opportunity, a detailed set of 30,000 household interviews, for a discussion of these data, see Boskin (1973) or Hall (1973) First, we estimate the expected real hourly wage rate facing each individual as a function of his personal characteristics This formulation is suggested by recent advances in the hedonic method of price measurement [6] We employ here an extended version of the particular type of wage equation proposed by Hall (1973), that is, we adopt an analysis-of-variance regression model,

$$\log(w/p) = \gamma' Z, \tag{8}$$

where the Z's represent personal characteristics such as race, sex, age, location, education, health, union membership, occupation, family status, etc We run separate regressions for each race-sex-occupation group, thereby allowing a complete set of interactions between these and all other effects In addition, we allow interactions between union membership and location, age and education, etc

These results are not without interest in themselves, the interested reader should consult Boskin (1972) and Hall (1973) For our purposes, however, they are important because they give us, for each individual, a method of estimating the wage he/she faces in each of 11 broad occupational classes In addition, we can estimate how that wage rate varies with age We thus can predict the course of potential lifetime earnings for each individual in each of several occupations he might choose to enter Of course, several refinements have to be made. We must at least attempt to allow for productivity growth, current 20-year-old workers will be working with an improved technology when they are 40, so we must estimate this productivity growth and adjust the wages of current 40-year-old workers accordingly to estimate wage rates facing 20-year-old workers 20 years from now. We estimate a constant rate of productivity growth from the average annual rate for 1960–70 We also assume that future potential earnings are discounted at a 5 percent rate

---

[6] See Hall (1973)

TABLE 1

SAMPLE CALCULATION OF THE PRESENT VALUE OF FULL-TIME EARNINGS BY OCCUPATION FOR A REPRESENTATIVE INDIVIDUAL*

Occupation	Present Value of Full-Time Earnings (Thousands of Dollars)
Professional/technical	134
Farmer	76
Manager	141
Clerical	99
Sales	97
Craftsman	109
Operative	95
Private household	67
Service	86
Farm laborer	59
Laborer	83

* White male high school graduate age 40

of interest, modest variations in the discount rate do not affect the results [7]

We thus estimate the present value of expected lifetime potential earnings as

$$PV_t = \sum_{t=t_0}^{65} \frac{W_t H_t (1 + \gamma)^{t-t_0+1/2}}{(1 + r)^{t-t_0+1}}, \tag{9}$$

where PV_t = the present value of potential work time remaining between age at t_0 and age 65, W_t = expected wage rate in year t, H_t = hours available for work in year t (= 2,000), γ = expected rate of productivity increase, assumed to be constant at the average annual rate from 1960 to 1970, and r = discount rate, assumed to be 5 percent

The second variable used in the study is the ratio of an index of training costs to current net worth [8] We use the data derived by Scoville (1966) on specific training requirements by occupation Out-of-pocket expenses are assumed to be one-third of foregone earnings Foregone earnings are computed by multiplying the present wage rate by necessary work time foregone in retraining We then take the ratio of training cost to net worth, the assumption being that the worker finances his retooling out of his own resources

The final variable we examine in this study is expected lifetime earnings lost due to unemployment If time spent unemployed were completely unproductive, that is, contained no element of leisure or investment in search activity, the worker would be indifferent, *ceteris paribus*, between two occupations, one offering $1 more in the present value of future full-time earnings, the other offering \$1 less in expected earnings foregone due to unemployment We could then subtract expected earnings foregone due to unemployment (net of unemployment insurance) from expected full-time earnings and use expected wage income as the focus of study. If, however, the measured unemployment includes a component of leisure or search, the time spent unemployed is valuable and the worker will require less than a \$1 decrease in relative expected earnings foregone due to unemployment to be indifferent to an occupation with a \$1 larger lifetime full-time earnings potential We have therefore separated these two components of expected income in order to attempt to test this hypothesis We estimate the expected duration of unemployment in a manner analogous to our procedure to estimate wages [9] We estimate analysis-of-variance hedonic unemployment equations of the Hall type (see Hall 1970) by regressing time unemployed on a set of personal characteristics We thus get an estimate of the expected unemployment facing a potential worker in each occupation and how this unemployment varies by age Following the procedure described above for potential earnings, we estimate the present value of potential earnings lost due to unemployment by using a formula similar to (7) in all but two respects We replace H_t, hours available for work in year t, with u_t, expected hours lost due to unemployment, and we replace w_t, the wage, by $(w_t - I_t)$, the wage net of the hourly equivalent of unemployment insurance [10]

---

[7] A sample calculation for a representative individual is presented in table 1

[8] Net worth is defined as cashable net worth, including the value of consumer durables and excluding the value of human capital

V. Empirical Results

Tables 2 and 3 present the empirical results, disaggregated by race and sex Table 2 presents occupation characteristics such as discounted present value of potential future earnings, discounted present value of expected lifetime earnings foregone due to unemployment, and our estimate of the ratio of training costs to net worth Since real wage and unemployment rates, even within race-sex groups (holding other variables, such as education, constant), differ substantially from occupation to occupation,[11] the substantial variation in our variables is hardly surprising These variables summarize the information (potentially) available to the worker in choosing an occupation [12]

---

[9] 1966 was a year of relatively full employment Projecting unemployment over the life cycle based on these data is the most reasonable procedure available to us but could result in misestimation, as relative unemployment by occupation varies over the business cycle

[10] Unemployment insurance is based on 1966 figures and is assumed to grow at the rate for the 1960–70 period

[11] See Reder (1955) for a discussion of these differentials

[12] We do not claim that they are the only ones which conceivably could affect occupational choice, rather, we assume that the influence of other variables, e g., those measuring tastes for certain types of work, is small relative to the variables considered here

TABLE 2
OCCUPATION CHARACTERISTICS FOR POPULATION SUBGROUPS (THOUSANDS OF DOLLARS)

	Mean for Adopted Occupations	Standard Deviation for Adopted Occupations	Mean Deviation for Nonadopted Occupations (Nonadopted – Adopted)
	Present Value of Lifetime Whole Earnings		
Total	63	56	3.6
White males	97	63	0.7
Black males	42	64	1.4
White females	52	48	–5.1
Black females	16	26	–12.8
	Ratio of Estimated Training Costs to Net Worth		
Total	0.20	1.8	0.25
White males	0.38	2.9	0.08
Black males	0.19	1.6	0.70
White females	0.19	1.6	0.28
Black females	0.13	1.5	1.02
	Present Value of Lifetime Earnings Lost Due to Unemployment		
Total	1.2	3.6	0.16
White males	1.2	2.0	0.15
Black males	2.6	8.2	–1.60
White females	0.8	6.0	–1.93
Black females	0.3	8.6	–5.28

SOURCE.—Computed from 1967 Survey of Economic Opportunity

TABLE 3
RELATIVE WEIGHTS OF POTENTIAL EARNINGS, TRAINING COSTS, AND FOREGONE EARNINGS DUE TO UNEMPLOYMENT IN THE CONDITIONAL LOGIT DECISION MODEL

Estimate No., Population Group, and Variables Included	Estimate of θ ($\times 10^{-2}$)	Standard Deviation* ($\times 10^{-2}$)	Likelihood L ($\times 10^{2}$)
1) White males			
Present value of potential earnings	1.084	0.075	–54.15
Training cost/net worth	–0.001	0.001	
Present value of time unemployed	–0.051	0.090	
2) Black males:			
Present value of potential earnings	0.072	0.013	–44.64
Training cost/net worth	–0.010	0.001	
Present value of time unemployed	–4.35	0.567	
3) White females			
Present value of potential earnings	0.875	0.076	–47.07
Training cost/net worth	–0.005	0.002	
Present value of time unemployed	–18.74	1.609	
4) Black females.			
Present value of potential earnings	0.378	0.132	–33.40
Training cost/net worth	–0.012	0.002	
Present value of time unemployed	–20.78	1.82	

SOURCE.—Based on 2,500 randomly selected observations from each population group from the 1967 Survey of Economic Opportunity

* Convergence was obtained after approximately 20 iterations, standard deviations reported are from the last iteration

The data offer some interesting insights into the workings of labor markets. The most obvious is the fact that opportunities and outcomes vary markedly among individuals—both within and across population subgroups—as evidenced by the large standard deviation of all variables. It is also interesting to draw some inferences across population subgroups Relative to other population subgroups, white males enter occupations with far higher training costs (the higher foregone earnings included in the numerator are mostly offset by larger net worth in the denominator) In addition, the mean deviation for the nonadopted occupations for all three occupation characteristics is much smaller for white males than for the other groups Given age, education, and the like, occupation is thus of less relative importance for white males than the other groups Finally, we note that blacks (male and female) have a much larger mean deviation of the ratio of training costs to net worth, this implies that the nonadopted occupations are relatively more expensive for blacks than for whites

Table 3 presents the evidence on the relative weights given to earnings, unemployment costs, and training costs The results for all population subgroups confirm our a priori expectations from human capital theory that (1) workers will tend to choose those occupations with the highest discounted present value of potential future earnings; (2) workers will tend to choose those occupations where retraining costs, in relation to net worth, are lowest, and (3) workers will tend to choose those occupations where, *ceteris paribus*, the discounted present value of expected earnings foregone due to unemployment is lowest [13]

All effects have the expected sign and almost all are measured quite precisely [14]

The most striking result is that white males tend to weight training costs and expected income foregone due to unemployment relative to expected full-time earnings much less heavily than the other groups This is consistent with the hypotheses of differential access to financing of training and education costs and of differential risk aversion

It is instructive to examine the ratio of the coefficients for expected full-time earnings and expected earnings foregone due to unemployment This figure varies markedly by race and sex The expected earnings foregone due to unemployment do not appear to exert much of an influence on white males The price ratio of trading $1 in full-time earnings for $20 in decreased earnings foregone due to unemployment suggests (perhaps) that measured unemployment for white males includes a not insignificant amount of voluntarily enjoyed productive activity—for example, leisure or job search. For the other population subgroups—females, many of whom work part-time and/or on and off throughout their lifetime, and black males—the price ratio is reversed. For example, white females appear to be willing to trade $1 less in foregone earnings due to unemployment for $20 of full-time earnings. This strongly hints of risk aversion or nonpecuniary costs involved in white female unemployment

---

[13] The predicted probabilities for adopted occupations ranged up to three times the predicted probabilities if random behavior was observed (probabilities equal to the percentages of jobs in each occupation).

[14] The results are similar when the training cost variable is left out of the equation These results are available upon request from the author It also should be pointed out that the likelihood ratio method may be used to test hypotheses about the coefficients

VI. Summary and Conclusion

We have applied the conditional logit decision model to the choice of occupation by individual workers to test some important implications of the theory of human capital The empirical results are quite consistent with the human capital hypothesis that workers choose occupations to maximize the discounted present value of potential lifetime work time Allowing for imperfect capital markets by including training costs relative to wealth and for unemployment by including the discounted present value of expected earnings foregone due to unemployment also yielded results consistent with a priori expectations

The conclusions apply to all four major race-sex population subgroups The apparent differences among subgroups are consistent with well-known labor market phenomena

References

Becker, G *Human Capital* New York· Nat Bur Econ Res., 1964.

Block, H , and Marschak, J "Random Orderings and Stochastic Theories of Responses " In *Contributions to Probability and Statistics Essays in Honor of Harold Hotelling*, edited by I Olkin et al Stanford, Calif Stanford Univ. Press, 1960

Boskin, M. "Unions and Relative Real Wages " *A.E.R* 62 (June 1972) 466–72.

———. "The Economics of the Labor Supply " In *Labor Supply and Income Maintenance*, edited by G Cain and H Watts. Chicago Markham, 1973

Fleisher, B. *Labor Economics Theory and Evidence*. Englewood Cliffs, N J.. Prentice-Hall, 1970.

Hall, R "Why Is the Unemployment Rate So High at Full Employment?" *Brookings Papers Econ Activity* 3 (1970) 369–402.

———. "Wages, Income, and Hours of Work in the U.S. Labor Force." In *Labor Supply and Income Maintenance*, edited by G. Cain and H. Watts. Chicago: Markham, 1973

McFadden, D *The Revealed Preferences of a Government Bureaucracy* Economic Growth Project, Technical Report no. 17. Berkeley Univ. California, 1968

Radner, R., and Miller, L "Demand and Supply in U.S. Higher Education A Progress Report." *A.E R.* 60 (May 1970) 326–40.

Reder, M. "The Theory of Occupational Wage Differentials." *A.E.R.* 45 (December 1955): 833–52

Schultz, T., ed. "Reflections on Investment in Man." *J.P.E.* 70, no. 5, suppl. (October 1962): 1–8.

Scoville, J. "Education and Training Requirements for Occupations." *Rev. Econ. and Statis* 48 (November 1966): 387–94.

Theil, H. "A Multinominal Extension of the Linear Logit Model." *Internat Econ. Rev.* 10 (October 1969): 251–59

Occupational Licensing and the Public Interest

Alex Maurizi

Office of Research, Assistant Secretary for Policy, Evaluation and Research, U.S. Department of Labor

Introduction

Occupational licensing has been justifiable in the view of legislatures on the grounds that it protects the public interest, often, however, it is the producers of the good or service who present this argument to the state legislatures This is hardly surprising, since the typical consumer is likely to suffer too small a wealth loss (in the form of higher prices) in the licensing of one more occupation, as has already been pointed out by others (Friedman 1962, p 143) The end result is the promotion of the interests of the producer group rather than those of the public Though this is what we economists would expect to happen, the principal difficulty in making progress with our case has been in trying to convince legislatures that the occupations receiving licenses do use their power to promote their own wealth at the expense of the consumer It is the purpose of this paper to present evidence on this point in order that the ground rules for discussing the problem with state legislators can be altered.

It will first be necessary to present the theoretical apparatus with which I will inspect and evaluate the data It should be noted at the outset that some of the ingredients of my model were mentioned in one very interesting earlier study of the purpose of licensing In particular, Moore (1961, p 112) expressed the proposition that "the more demand is growing for the occupation, the greater the return from licensing." Moore concluded that the evidence indicated the tendency to license was not correlated with its monetary return (1961, p 113) I have formulated the hypothesis somewhat more precisely and have used appendix data from the same study utilized by Moore but which appendix data he neglected to analyze (Council of State Governments 1952, p 92), my investigation tends to support the same basic hypothesis which his investigation failed to support

A Simple Model of Licensing

The usual short-run and long-run supply curves are presented in figure 1 together with a market demand curve; we assume the initial position to be one of short-run competitive equilibrium Licensing initially raises

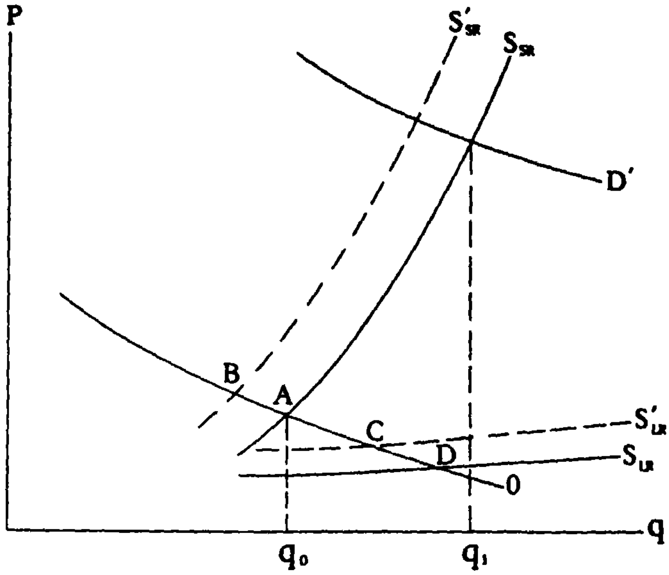


Fig 1

entry costs by imposing at least a fee and usually also training requirements,[1] and shifts the short-run and long-run supply curves upward and to the left. Under these conditions, after licensing is imposed there will be a drop in entrants and an income gain distributed to members of the profession. Once the magnitude of these long-run gains are realized and become widely known, there will be an increase in entrants once again In the absence of other changes, prices (incomes) will decline in the long run to position *C* The entry restriction will prevent position *D* from being reached. Arriving at this long-run equilibrium position may take considerable time, during the early years of which period, of course, established practitioners are enjoying higher incomes

---

[1] In addition to requiring payment of fees (usually nominal, but amounting to $1,500 for glaziers in Chicago and $500 for truckers in Seattle in 1955) and completion of formal training in approved institutions, there are numerous ways in which the licensing agency or body may increase the cost and inconvenience of attempting to enter the occupation For example, by requiring that examinations be conducted at only one location in the state, the agency increases travel costs for some applicants By sometimes requiring applicants to bring all their own tools and equipment to the examining station, the agency adds a considerable cost to some applicants Some cosmetologists, for example, must bring their own human model, some truck drivers must bring their own trucks Furthermore, in some occupations knowledge about what applicants must do to pass the examination is difficult to obtain, plumbers' examinations are so unpredictable that applicants often must take the exam once (figuring to fail it) in order to find out what it is like. And, of course, most occupations require applicants to be of "good moral character", even the (otherwise) best qualified applicants have been turned down on these grounds sometimes, as the experience of one top-ranking law school graduate who refused to sign a loyalty oath testifies For details regarding these and numerous other similar cases, see Gellhorn (1968, pp 105–52) For simplicity, I shall summarize these various elements of cost to the applicant under the category of "fees and training requirements." There is a distinction to be made, to be sure, between the effects of requiring an exam and charging a fee for examinees and the establishment of training requirements The former usually are nominal requirements, though the above discussion clearly indicates there are exceptions Though some people are probably turned away even by these requirements, their numbers are probably negligible in comparison to the numbers turned away by the formal training requirements In most occupations applicants are required not only to complete a specified number of years of high school or college education, but also to attend an approved school designed to impart the necessary skills Though I shall speak of "fees and training requirements" in connection with licensing, it should be understood that the principal entry restriction occurs as a result of required training.

Increases in fees and training requirements do occur over time in an occupation, of course, one need only compare the length of training required to become a physician today with that required less than a century ago, "The training legally required of medical practitioners has thus been extended since the beginning of the century from three or four years of professional education preceded by one or two years of high school to six or seven years preceded by graduation from high school" (Friedman and Kuznets 1954, p 10) Changes of this nature occur slowly over time and are not likely to be a major instrument through which licensed occupations make alterations in their wealth position from year to year For this latter purpose it is more likely that licensed occupations will utilize the pass rate on the yearly examinations The extent to which the pass rate is altered will depend on, among other things, the extent of excess demand or supply existing at the time The greater the extent of excess demand[2] generated by a shift in *D* (initially evidenced by actual,

---

[2] My discussion will be in terms of a positive excess demand, though the same model applies to a negative excess demand (excess supply) Hicks has pointed out that economic expansion followed by stagnation provided the historical early trading centers around 1400 with an even greater incentive to enter into arrangements in restraint of competition "The growth of guilds and similar associations is well known." "As opportunities in general close in, or seem to close in, the fields in which it becomes tempting to protect oneself by agreements with one's competitors become more extensive " See Hicks (1969, p 57) More recently, the depression of the 1930s had its impact on the admission of students into medical school Following annual increases from 1929, from 1934 to 1939 there was a decline in the number of and percentage of applicants accepted A comment made then by the New York State Board of Medical Examiners' secretary is of interest here "The trend is still in an early stage, but in law, medicine, dentistry, and other professions under control of state licensure, the signs are apparent Without intention or design, the far-reaching steps taken by the physicians to raise educational standards during the past twenty-five years has resulted in limiting the number of students. Now [1933], realizing the advantages of this unplanned restriction, leaders are taking definite steps to cut down the professional class" (Friedman and Kuznets 1954, pp 12–14). If a decline in the number of applicants should prove inadequate during periods of large excess supply, the licensing agency must seek an alternative instrument It is likely at this point more thought will be devoted to identifying and eliminating "quacks" or incompetents. During the 1930s in medicine, graduates of foreign medical schools faced increasingly difficult entry requirements (Friedman and Kuznets 1954, pp. 17–20) At some point the licensing agency might even offer practitioners in the area of excess supply a monetary incentive to leave. Though I could find no evidence of this happening, it is interesting to note that labor unions sometimes pay some of their members to cease seeking employment in the industry. The Amalgamated Clothing Workers at the Chicago plant

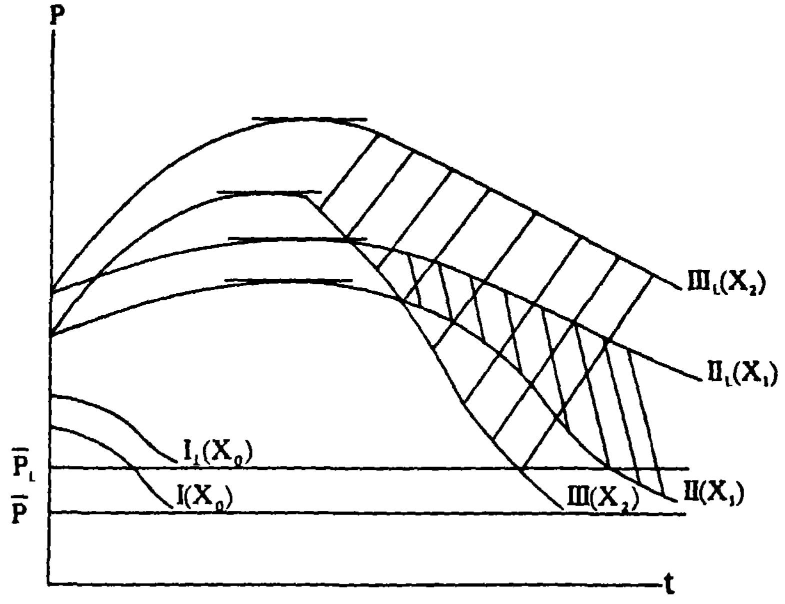


FIG. 2

temporary shortages at initially prevailing prices), the more rapidly will prices (hence incomes) rise, the more rapidly will entry rise, and thus the greater will be the incentive for the licensing agency to lower the pass rate[3] The reasoning here may best be illustrated with the aid of figure 2

The nature of the price adjustment process[4] is described in figure 2 Curves I (X_0) and I_L (X_0) assume no excess demand and a short-run equilibrium position initially above the long-run position; with no shifts in D or S, price will eventually fall to its long-run level, higher for the

---

of Hart, Schaffner, and Marx had experienced a change in cutting methods which created an excess supply of cutters. The union took $25,000 from their unemployment insurance fund to help compensate any dismissed worker, who took the money with "the understanding that he was to leave the industry altogether and go into some other line of work." See "American Trade Unions and the Problem of Unemployment" (1928, p 483), see also Slichter (1941, p 131) I am indebted to Melvin W Reder for this reference.

[3] The relation between the extent of aggregate excess demand and the rate of change of labor earnings (wage rates) has been investigated empirically by Lipsey in relation to the Phillips curve, of course. Using the inverse of the unemployment rate as an indicator of excess demand, Lipsey (1960) found that wage rate changes from 1862 to 1957 were well explained by the level of excess demand and the rate of change of the level of excess demand.

[4] See Reder (1960) for a discussion of the adjustment process in connection with the union influence on wages: "The ethics of the queue affect the adjustment process of labor markets, but only rarely their equilibrium values." As Reder points out, to raise the long-run equilibrium price level would require entry restriction activity on the part of the union or, in our case, licensed occupations (see Reder 1960, pp. 351–54).

licensed occupation than the nonlicensed one Curve II (X_1) begins at time zero with the short-run equilibrium value of P which is lower than for the licensed occupation indicated by II_L (X_1); both are above their long-run equilibrium values Curve II (X_1) corresponds to the price response to a small amount of excess demand ($\dot{P}$, the slope of the curve, is initially relatively low) in a competitive situation with unrestricted entry, while III (X_2) corresponds to the price response to a larger amount of excess demand with unrestricted entry. Note that III (X_2) rises faster to a higher peak value and then falls to $\overline{P}$ faster than II (X_1); this rapid falling off, we suppose, is due to the strength of the signal to potential entrants generated by the rapid price increase [5] In a competitive situation with entry restriction, however, the price will rise to a higher peak value than II (X_1) (given the same excess demand) and will fall off much more slowly, as indicated by II_L (X_1) But with a greater excess demand ($\dot{P}$ is greater) entry will be restricted more, prices will rise to a higher peak value and trail off more slowly, as indicated by III_L (X_2) compared to II_L (X_1) Thus, even if the increased entry restrictions do not alter the long-run equilibrium price level, there is a substantial difference in the nature of the adjustment process [6]

Our primary task now is to measure the empirical counterparts of these two theoretical variables.

The Data

Though several studies have collected data on licensed occupations by state (Council of State Governments 1952, 1968, Vaughan 1963; Shimberg

---

[5] Alchian has recently pointed out that even in a competitive situation, downward shifts in demand would cause downward adjustments in price, notably the price of labor services, to occur with a considerable lag Workers must discover by searching that such a demand shift has occurred Obviously, the greater the downward shift, the sooner the price adjustment occurs Once the price shift has occurred, it will in itself serve to communicate what has happened In this study I am concerned with the change in price itself once it has occurred. The same principle is at work, however I assume that information about the permanence of any shift in demand is costly and the *extent* of the price rise is one indicator of this permanence to potential entrants See Alchian (1969)

[6] If we suppose that the demand and supply curves are linear, then the following is true $D = -ap + c$, $S = bp + d$, $p = k(D - S) = kX$, where $X = D - S = -ap + c - (bp + d) = -(a + b)p + (c - d)$, $dX/dt = -(a + b)\dot{p}$ Let p^* be the short-run equilibrium price and p the actual price. Then p^* satisfies $X = 0 = -(a + b)p^* + (c - d)$. Multiplying by k

$$0 = -k(a + b)p^* + k(c - d). \qquad (1)$$

But since $\dot{p} = kX$,

$$p = -k(a + b)(p^* - p) \qquad (2)$$

That is, the rate of increase in prices will initially be larger, the larger the shift in D; as p approaches p^*, the rate of increase in p will gradually decrease until it reaches zero when the new short-run equilibrium price is reached. See Arrow and Capron (1959) Thus, the peaks of the curves drawn in fig 2 correspond to the positions of new short-run equilibrium where $\dot{p}$ is momentarily zero. It should be pointed out that the adjustment process from the new short-run equilibrium position to the final long-run equilibrium position has not been previously investigated.

1968; Greene 1969, Shimberg, Esser, and Kruger 1973), only one previous study (Council of State Governments 1952) has collected and published the kind of data needed for our study. For approximately 24 occupations licensed in various states, the Council of State Governments collected and published by state the number of practitioners licensed in the state, the number of applicants for licensure, and the number of new licenses granted (1952, appendix tables 6, 7) for 1940 and 1950. For each occupation in each year I chose the number of states for which data were available for all three of these concepts. Most of the occupations chosen were licensed in nearly all of the states, but the data were usually available for only a small fraction of them The data do present some difficulties, however First, there is seldom uniform agreement among the statutes regarding the definition of what constitutes a licensed occupation, so that some caution must be used in interpreting the results Second, the number of new licenses granted is sometimes greater than the number of new applicants This is because new licenses granted by reciprocity are included in the former category while applicants for licensure by reciprocity are not included in the latter I attempt to make an adjustment for this possible bias in the data in my study

The Variables of the Hypothesis

As an indicator, X, of excess demand, I choose the ratio of new applicants in the state to the number of licensed practitioners already in the state The greater this ratio, the greater I assume excess demand for that occupation in the state to be in that year Using figure 1 we can see that X is essentially $(q_1 - q_0)/q_0$ Given q_0, obviously the greater is q_1, the greater is X It seems quite reasonable to suppose that the number of new people who would like to enter the occupation (and indicate as much by applying for a license) is a good approximation to $(q_1 - q_0)$

As an indicator of the leftward shift in the supply curve caused by entry restriction, I choose the pass rate, P, or the ratio of the number of licenses granted to the number of applicants [7] The pass rate is clearly a variable under the control of the licensing board Decisions regarding the pass rate normally would be made after all the applications had been received. If the license exams are held, say, once a year, then the board members clearly have knowledge of X before the exams are given The same holds if the exams are given periodically during the year, except that applicant rates for shorter periods must be examined for possible trends in yearly data. Since in most cases the licensing board is not required even to declare the reasons for failure, it is clearly possible for the board to alter P in response to X.

---

[7] In a previous study it was assumed that P was an exogenous variable, though it was recognized that this assumption might be faulty In this study P is not exogenous, it is the variable whose behavior we wish to explain. See Benham, Maurizi, and Reder (1968)

My hypothesis can now be stated as follows

$$P = A + BX, \qquad B < 0 \tag{3}$$

(To test this hypothesis I ran regressions for each occupation by state for 1940 and 1950.)

As mentioned above, the data are such that my estimate of P is greater than one for some observations for which P is actually less than one. To correct for this I have run the regressions with estimated P's greater than one eliminated.

Although the number of observations eliminated in this way was quite small (only 12 of 238 observations were eliminated in 1940, only 29 of 370 in 1950, 17 of which were among pharmacists), I investigated the possible bias introduced by running the regressions without eliminating these observations. The results were not substantially altered.[8]

The Results

The results of my investigation are reported for 1940 in table 1 and for 1950 in table 2. Regression coefficients with their t-values in parentheses below are recorded, as well as corrected R^2 and the degrees of freedom. For each regression I have also calculated and recorded the elasticity of P with respect to X evaluated at the point of means.

The reader should note that of 36 cases, A is significantly positive in all 36 and B is negative for 27 of them, and of these B is significant 17 times. The amount of variation explained in most of the significant cases is quite high for a simple bivariate cross-section relationship, it is over .1 in 16 of the significant cases, over .2 in 12 of the significant cases, and over .3 in nine of the 17 significant cases. The elasticity is less than -1.0 in all but one of the significant cases and varies primarily between -1.0

---

[8] The bivariate regression coefficient remained negative and at least as significant in nine out of 10 cases, actually increasing negatively in six of these nine for 1950, in one case the sign changed from an insignificant positive to an insignificant negative one. The $\bar{R}^2$ did drop somewhat in two of these nine cases (from .28 to .17 and from .28 to .21) and quite substantially in one (from .17 to .04), that one being pharmacists, for which the number of observations was nearly doubled. The drop in $\bar{R}^2$ in the other six cases was slight. The results for 1940 were similar. The bivariate regression coefficient became larger negatively and remained as significant in six of 11 cases, in one case the negative coefficient became significant; in three of 11 cases the sign of the coefficient went from (insignificant) positive to (insignificant) negative, in one case (physicians) it went from significant negative to insignificant positive. The $\bar{R}^2$ was only slightly less in all but this one case, where it decreased from .61 to .05. In other words, except for physicians in 1940 and pharmacists in 1950, excluding all observations where $P > 1$ does not alter the results.

TABLE 1
$P = A + BX$ 1940

Occupation	A $T(A)$	B $T(B)$	df	$\bar{R}^2$	E
01 Accountants	0.521 (3.51)	−0.844 (−1.39)*	12	.07	−.52
02. Architects	0.766 (5.23)	−0.665 (−2.38)**	9	.32	−.31
03 Attorneys	1.225 (7.70)	−8.598 (−3.04)***	9	.45	−.69
04 Barbers	0.776 (6.88)	1.357 (0.38)	12	.00	.05
05. Beauticians	1.023 (27.65)	−0.896 (−7.63)***	10	.84	−.20
06 Chiropodists	0.558 (3.96)	2.334 (1.70)*	13	.12	.23
07 Chiropractors	0.940 (24.48)	−0.056 (−0.49)	9	.00	−.01
08 Dentists	1.016 (16.62)	−5.305 (−3.75)***	22	.36	−.26
09 Embalmers	0.837 (15.87)	−0.847 (−4.36)***	11	.60	−.13
10 Engineers, professional	0.655 (9.98)	0.998 (1.46)	15	.07	.11
11 Funeral directors	1.067 (6.05)	−5.488 (1.76)	3	.34	−.33
12 Nurses, registered	0.906 (8.59)	−0.205 (−0.17)	11	.00	−.02
13 Optometrists	0.877 (5.53)	−2.469 (−1.31)	12	.05	−.30
14 Osteopaths	0.850 (10.79)	0.091 (0.29)	14	.00	.01
15 Pharmacists	0.829 (15.46)	−0.049 (−0.22)	11	.00	−.01
16 Physicians	1.002 (21.17)	−0.611 (−5.28)***	16	.61	−.15
17 Real estate brokers and salesmen	1.017 (15.81)	−0.828 (−7.30)***	5	.90	−.47
18 Veterinarians	0.928 (21.18)	0.047 (0.34)	9	.00	.01
All occupations combined	0.83 (43.50)***	−0.460 (−6.10)***	237	.13	−.08

* Significant at the 10% level
** Significant at the 5% level
*** Significant at the 1% level

and −0.50. That is, a 10 percent increase in excess demand will generate a fall in the pass rate of between 10 percent and 5 percent.[9] In general,

---

[9] Since my dependent variable is of the form x/y, and my independent variable, of the form y/z, the questions arise as to whether there is spurious correlation present and if so whether it tends to generate a negative relationship. It should be noted, however, that the question of spurious correlation generally arises when a hypothesis pertains to a set of undeflated variables and the data are adjusted by dividing through by another variable. Thus, if an initial hypothesis suggests $x = f(y, z)$ and the data for x, y, and z are all divided by w so that the regression run is of the form $x/w = f(y/w, z/w)$, spurious correlation may arise. See Kuh and Meyer (1955) for a detailed discussion. To the best of my knowledge, the extent of spurious correlation when the regression run is of the form $x/y = f(y/z)$ has not been investigated. Therefore, I have made a simulation test to determine the extent of any negative bias, if any. The values for x, y, and z were picked

TABLE 2
$P = A + BX$, 1950

Occupation	A $T(A)$	B. $T(B)$	df	$\bar{R}^2$	E
01. Accountants	0.486 (6.45)	−0.329 (−2.63)***	19	.23	−.49
02. Architects	0.713 (9.79)	−0.139 (−1.04)	22	.00	−.74
03. Attorneys	0.662 (7.50)	−0.098 (0.52)	5	.00	.04
04. Barbers	0.860 (11.31)	−1.251 (−1.08)	22	.01	−.10
05. Beauticians	0.916 (11.12)	−0.909 (−0.86)	14	.00	−.07
06. Chiropodists	0.917 (9.56)	−0.557 (−0.72)	19	.00	−.07
07. Chiropractors	0.822 (11.33)	−0.135 (−2.08)**	17	.16	−.06
08. Dentists	0.955 (19.55)	−1.650 (−2.36)**	26	.14	−.12
09. Embalmers	0.839 (12.05)	−0.881 (−2.87)***	18	.28	−.13
10. Engineers, professional	0.693 (16.17)	−0.001 (−0.01)	25	.00	−.00
11. Funeral directors	0.866 (8.31)	−0.749 (−2.23)**	9	.28	−.13
12. Nurses, registered	0.786 (7.08)	−0.023 (0.02)	16	.00	−.00
13. Optometrists	0.853 (11.19)	−1.095 (−2.49)***	25	.17	−.24
14. Osteopaths	0.731 (7.15)	0.347 (0.717)	17	.00	.04
15. Pharmacists	0.849 (17.30)	−0.515 (−2.60)***	28	.17	−.09
16. Physicians	0.890 (23.45)	0.131 (0.67)	17	.00	.01
17. Real estate brokers and salesmen	0.938 (12.99)	−0.636 (−5.81)***	19	.62	−.54
18. Veterinarians	0.917 (18.49)	−0.028 (−0.15)	17	.00	−.00
All occupations combined	0.80 (54.10)***	−0.29 (−7.90)***	369	.14	−.07

** Significant at the 5% level
*** Significant at the 1% level

---

at random from a normal distribution with unit mean and variance. This choice was made because the data indicate that the coefficient of variation (the ratio of the standard deviation to the mean) for each of the three variables varies around unity, table 3 presents the coefficients of variation for 1940 only, since the results for 1950 are similar. As is clear from table 3, the variation in Y is sometimes greater than that in X or Z. In six cases the variation in Y is greater, but in seven cases it is less than that in either X or Z, in the remaining cases it is in between and nearly the same as that for either X or Z. The numbers in all three cases appear to be scattered above and below, but generally close to unity. The simulation run consisted of 20 regressions whose sample size varied from 34 to 64. The mean regression coefficient equaled −.027; the t-test on this mean equaled −3.07, significant at the 5 percent level. In other words, there is a negative bias present in the regression coefficient. The size of the bias, however, is no more than one-fifth to one-tenth, often only one-hundredth, the size of the regression coefficients presented in the tables. The bias will cause an overestimation of the effect we are measuring but is not likely to cause a change in the sign of the coefficient itself.

TABLE 3
COEFFICIENTS OF VARIATION, 1940

Occupation	No Passing X	No Applying Y	Stock Z
01	0.96	1.08	0.93
02	1.32	1.06	1.17
03	0.63	0.76	0.71
04	0.89	0.89	0.78
05	1.05	1.05	1.08
06	2.79	2.85	1.99
07	2.19	2.17	1.86
08	1.02	0.96	1.10
09	1.05	1.08	0.95
10	0.82	0.76	1.03
11	0.66	0.72	0.53
12	1.24	1.26	0.93
13	1.19	1.00	0.95
14	1.23	1.22	1.33
15	0.81	0.72	0.80
16	3.21	1.38	1.34
17	0.80	1.94	1.14
18	0.80	0.77	0.91

for significant cases in 1950 the corrected R^2 is less than in 1940 With only two points it is difficult to determine a trend, but if these points are indicative of such a trend, it would indicate that my model has less explanatory power in 1950 That may be due to changes in some missing explanatory variable, one obvious candidate for inclusion is the level of practitioner incomes

The addition of practitioner incomes along with the applicant stock ratio as independent variables raises some question regarding the possibility of multicollinearity Though the income variable pertains to the *level* of income while the applicant-stock ratio is derived from and is an indicator of the *rate of rise* in the level of income, there is nonetheless a question as to whether high-income states and rising-income states are coincident In my pooled sample, the correlation matrix indicates mostly a small, positive correlation between the two variables, the zero-order correlations equal 12, 38, 19, − 02, 05, − 16, and 01, respectively, for the occupations in my pooled sample The problem of multicollinearity does appear to be of minor importance in this case

If in two states an occupation is experiencing the same increase in excess demand, but the average income of practitioners in one state is higher, then the capitalized gain from entry restriction is greater in the higher-income state. I conclude that the tendency to restrict would therefore be greater in the higher-income state. In order to test this hypothesis I have introduced one additional explanatory variable, Y, the mean income of practitioners in the state in the years 1940 and 1950. My hypothesis now is $P = A + BX + CY, B < 0; C < 0$ Since mean incomes for some occupations are not available in these years, my sample of occupations is considerably reduced. Of the original 18 occupations only eight are left in 1940 and 10 in 1950 The regression results are reported in table 4 for 1940 and table 5 for 1950.

Of the eight cases in 1940, X is significantly, negatively related to P four times and positively but not significantly related to P four times The variable Y is negatively related to P five times and significantly twice, while being positively related to P three times, once significantly Corrected R^2 is increased in only two cases by adding Y The elasticity estimates with respect to X are quite similar to those reported in table 1 except for attorneys. The result in table 4 indicates a very large, negative elasticity of -5 87 for attorneys in 1940 with Y present, compared to -0 69 in 1940 with Y absent. Both variables are significantly, negatively related to P, and corrected R^2 is quite high (78) and significant This result would indicate that for two states in which attorneys were experiencing the same average incomes, the one in which the excess demand was 10 percent greater would exhibit a pass rate nearly 60 percent lower For all other occupations in 1940 (and 1950), this elasticity was less than -0 5, indicating that (holding average incomes constant) a 10 percent increase in excess demand would result in at most a 5 percent decrease in the pass rate The elasticity of the pass rate with respect to Y was less than -1 0 in 1940 (and 1950) For occupations in two states experiencing the same excess demand, the pass rate would fall by 10 percent (at most) for every 10 percent increase in average incomes

The results reported in table 5 for 1950 are similar to those in table 4 The variable X is negatively, significantly related to P in five out of 10 cases, while positively in five out of 10 cases but only once significantly The variable Y is negatively related to P in eight out of 10 cases, three of them significantly, while positively related twice, once significantly Corrected R^2 is increased in six out of 10 cases and decreased in only two cases The elasticity of P with respect to X is quite similar when Y is present The elasticity with respect to Y varies mainly between -0 1 and -0 5, as in 1940

The improvements from including Y are somewhat better in 1950 than in 1940 Since unemployment was much more widespread in 1940, it is understandable that reported incomes would not be an accurate indicator of permanent incomes and might therefore produce less noticeable results On the whole, however, it appears that introducing the variable Y was worthwhile The results are generally as expected

Finally, to enlarge the sample size I have pooled the results for 1940 and 1950 occupations whenever data are available for each The results are presented in table 6. The regression coefficient for the applicant-stock ratio is significant in only two of seven cases, but both those times it is negative. The R^2 for these two (dentists, and real estate brokers and salesmen) is .44 and .65, respectively, and highly significant. The size of the

TABLE 4

$P = A + BX + CY$ 1940

Occupation	A $T(A)$	B $T(B)$	C $T(C)$	df	$\bar{R}^2$	F	E_{PX}	E_{PY}
02. Architects	0.82 (2.16)**	−0.65 (−2.06)**	−0.000029 (−0.14)	8	.23	2.54	−0.30	−0.22
03 Attorneys	1.59 (11.47)***	−8.04 (−4.49)***	−0.000284 (−3.82)***	8	.78	18.93***	−5.87	−0.54
08. Dentists .	0.98 (16.65)***	−6.16 (−4.39)***	0.000091 (1.96)**	21	.43	9.84***	−0.31	0.075
10 Professional engineers	0.70 (1.78)**	0.0035 (0.37)	0.000012 (0.08)	14	.00	0.17	0.01	0.03
14 Osteopaths	0.71 (4.88)***	0.098 (0.31)	0.000400 (1.09)	13	.00	0.64	0.014	0.16
15 Pharmacists	1.42 (5.24)***	0.077 (0.38)	−0.000552 (−2.22)	10	.20	2.51	0.008	−0.67
16. Physicians	0.98 (17.27)***	0.043 (0.65)	−0.000012 (−0.19)	8	.00	0.28	0.005	−0.01
17. Real estate brokers and salesmen	1.11 (6.96)***	−0.76 (−7.38)***	−0.000143 (−0.99)	4	.90	29.31***	−0.41	−0.22

NOTE.—P = pass rate, X = excess demand, Y = mean income by state (1940 Census)
** Significant at 5% level
*** Significant at 1% level

TABLE 5
$P = A + BX + CY$ 1950

Occupation	A T(A)	B T(B)	C T(C)	df	$\bar{R}^2$	F	E_{PX}	E_{PY}
01 Accountants	1 03 (1 59)	−0 30 (−2 30)**	−0 00014 (−0 84)	17	.20	3 43	−0 44	−0 12
02 Architects	−3 65 (−2 26)	0 21 (1 74)*	0 00073 (2 56)**	3	59	4 62	0 12	0 76
03 Attorneys	0 18 (2 83)	0 11 (0 68)	−0 00018 (−1 78)*	4	21	1 78	0 02	−0 16
04 Barbers	1 33 (3 20)	−1 57 (1 38)*	−0 00019 (−1 15)	20	04	1 43	−0 16	−0 59
05. Beauticians	1 70 (4 10)	−2 23 (−2 03)**	0 00029 (−1 84)**	12	22	2 92	−0 16	−0 83
08 Dentists	1 31 (7 45)	−3 49 (−3 59)***	−0 00004 (−1 55)*	17	51	10 78***	−0 17	−0 31
10 Professional engineers	1 18 (2 25)	0 13 (0 17)	−0 00012 (−1 05)	18	00	0 58	0 02	−0 80
15 Pharmacists	1 13 (3 55)	0 18 (0 75)	0 00007 (−0 96)	19	00	0 61	0 02	−0 36
16 Physicians	0 51 (1 35)	0 30 (0 18)	0 00005 (1 18)	12	00	0 80	0 02	0 45
17. Real estate brokers and salesmen	0 93 (1 10)	−0 56 (−3 37)***	−0 00001 (−0 02)	13	46	7 34**	−0 38	−0.06

NOTE —P = pass rate, X = excess demand, Y = mean income by state (1950 Census)
* Significant at 10% level.
** Significant at 5% level
*** Significant at 1% level

TABLE 6
COMBINED 1940 AND 1950 RESULTS $P = A + BX + CY$

Occupation	A. $T(A)$	B $T(B)$	C· $T(C)$	df	$\bar{R}^2$	F	E_{PX}
2 Architects	0 682	−0 249	−0 00000	14	.00	0 79	−0 13
	(4 07)***	(−1 18)	(0.30)				
3 Attorneys	0 793	0 113	−0 00000	15	00	1 06	0 02
	(12 38)***	(0 73)	(−1 45)*				
8 Dentists	0 987	−4 787	0 00001	41	41	16 21***	−0.23
	(23 84)***	(−5 53)***	(2 34)***				
0 Professional engineers	0 858	0 006	−0 00005	35	03	1 60	0 01
	(8 04)***	(0 72)	(−1 63)**				
5 Pharmacists	0 835	0 022	−0 00000	32	00	0.01	000
	(13 66)***	(0 14)	(0 08)				
6 Physicians	0 965	0 038	−0 00001	23	02	0 72	0 00
	(21 80)***	(0 25)	(−1 18)				
7 Real estate brokers and salesmen	0 911	−0 627	0 00001	20	61	18 23***	−0 40
	(7 78)***	(−6 04)***	(0 26)				

NOTE —P = pass rate, X = (applicant stock ratio) excess demand indicator, Y = average income of practitioners n state Regression coefficients are reported with t-values in parentheses below
* Significant at 10% level
** Significant at 5% level
*** Significant at 1% level

regression coefficients in these two cases is approximately the same as before The income variable is negatively significant twice (for attorneys and professional engineers) and is positively significant once (for dentists) The elasticity estimate, E_{PX}, for dentists and for real estate brokers and salesmen is − 23 and − 40, respectively, similar to previous findings

Conclusion

With a rather simple model of the short-term behavior of licensing boards we are able to explain a substantial part of the pass rate of licensing boards across states for nearly half of all the cases studied in 1940 and 1950 The findings indicate that a 10 percent increase in excess demand generates a decrease in the pass rate varying primarily from 1 percent to 10 percent depending on the occupation, and that a 10 percent increase in average practitioner incomes produces up to a 10 percent decrease in the pass rate This evidence tends to confirm the notion that the power of licensing boards is often used to prolong the period of higher incomes resulting from increases in excess demand for the services of the occupation in question and that the instrument then used to accomplish this purpose is alteration of the pass rate on the licensing examination.

It is disturbing, however, that my model has no explanatory power at all in approximately half of all the cases considered. I have made only a first step in gathering evidence to support the notion that licensing of occupations is not in the public interest My hope is that this avenue of investigation can be pursued further when better, more reliable, and more comprehensive data become available. If this study speeds the accumulation, publication, and analysis of such data, it will have accomplished a great deal.

References

Alchian, A C "Information Costs, Pricing, and Resource Unemployment" *Western Econ J* (June 1969), pp. 109–28

"American Trade Unions and the Problem of Unemployment" *Monthly Labor Rev* 26, no. 3 (March 1928) 483

Arrow, K. J, and Capron, W M "Dynamic Shortages and Price Rises The Engineer-Scientist Case" *Q.J E* 73, no. 2 (May 1959) 292–308

Benham, L; Maurizi, A, and Reader, M W "Location, Migration, and Remuneration of Medical Personnel Physicians and Dentists" *Rev Econ and Statis*. 50, no 3 (August 1968) 332–47

Council of State Governments. *Occupational Licensing Legislation in the States* Chicago, 1952

——— *Occupations and Professions Licensed by the States, Puerto Rico, and the Virgin Islands* Chicago, 1968.

Friedman, M "Occupational Licensure" In *Capitalism and Freedom* Chicago. Univ Chicago Press, 1962

Friedman, M, and Kuznets, S *Income from Independent Professional Practice* New York Nat. Bur Econ Res, 1954

Gellhorn, Walter "The Right to Make a Living" In *Individual Freedom and Governmental Restraints* New York Greenwood, 1968

Greene, Karen *Occupational Licensing and the Supply of Nonprofessional Manpower* Washington Department of Labor, 1969.

Hicks, J R *A Theory of Economic History* London. Oxford Univ. Press, 1969

Kuh, Edwin, and Meyer, John "Correlation and Regression Estimates When the Data Are Ratios" *Econometrica* 23 (October 1955) 400–416.

Lipsey, Richard G "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the U K, 1862–1957." *Economica* 28, no 105 (1960). 3–31

Moore, Thomas G. "The Purpose of Licensing" *J Law and Econ* (October 1961), pp 93–117

Reder, M W "Job Scarcity and the Nature of Union Power" *Indus and Labor Relations Rev.* 13, no 3 (April 1960)· 349–62.

Shimberg, Benjamin. *A Pilot Study to Determine the Feasibility of Investigating Nationally the Impact of Licensing Practices on the Availability and Mobility of Nonprofessional Manpower in Occupations where Skill Shortages Exist* Princeton, N J. Educ Testing Service, 1968

Shimberg, B, Esser, B, and Kruger, D *Occupational Licensing Practices and Policies*. Washington Public Affairs Press, 1973.

Slichter, Summer H *Union Policies and Industrial Management*. Washington Brookings Inst., 1941.

Vaughan, M. R. *Occupational Licensing in Wisconsin* Madison· Wisconsin Legislative Reference Bur, 1963.

On Equalizing the Distribution of Political Income

Dennis C. Mueller
Cornell University

Robert D Tollison
Texas A & M University

Thomas D Willett
US Treasury

I. The Problem

Individuals obtain utility from the consumption of both public and private goods There has been considerable interest recently in how money income—the power to purchase private goods—should be distributed (Fair 1971, Thurow 1971) In this paper we explore the complementary question of how political income, or voting rights, should be distributed to obtain a desirable distribution of individuals' utility from public goods.

It is important to distinguish two concepts of equality in the discussion of public policy which we will employ to discuss the distribution of political income.[1] Ex ante equity refers to fairness in the sense that risks or opportunities before the fact are equalized Ex post equity is fairness in the sense of actual outcomes of social processes involving risk and uncertainty In most cases ex post equity is probably of greater importance than ex ante equity, although, as we shall discuss, equalitarians have curiously failed to extend considerations of equity in the distribution of political income to an ex post setting [2] There are, however, cases where ex ante equity would be more important For instance, this would be the case where equality of opportunity is the objective or where it may be more desirable to substitute a mechanism to achieve ex ante equity (or some combination of the two concepts) because ex post equity is impossible or

We are grateful to James M Buchanan for useful comments Any remaining errors are our responsibility.

[1] For an extended treatment of the two concepts of equity, see Pauly and Willett (1972) and the references cited there

[2] Note, however, the following quotation from a paper by Kolm (1971, p. 8). "Les Américains ne s'intéressent pas à la justice sociale, du moins au sens où l'entendons en France (Léon Walras écrivait déjà. 'La justice est une idée gréco-latine et française'). En particulier, la notion d'égalitarisme dans les résultats—plutôt que dans les chances—a beaucoup moins d'attraits pouveux que pour nous "

too costly to achieve. We will argue that the lack of consideration of establishing equality in the ex post distribution of political income cannot be explained by either the impossibility of achieving such an outcome or the high costs of pursuing such a goal

Political equality is generally thought to be more feasible and desirable than economic equality. Economic equality—for example, in Lerner's sense of complete ex post equality of utility distribution—is typically expected to generate serious excess burden and disincentive effects, since the generation of private income takes a large part of an individual's time, that is, the elasticity of supply is higher [3] Thus, emphasis in achieving greater equality of the private income distribution focuses on a mix of mechanisms which trade off some degree of ex post inequality in private incomes against the disincentive effects on the total amount of private income generated For example, mechanisms such as fair employment laws can be viewed as generating greater ex ante equity, or equality of opportunity, while a progressive income tax can be viewed as generating greater ex post equality

Discussion of the analogous problem of the distribution of political income has been characterized primarily by the discussion of generating mechanisms to achieve ex ante equality of opportunity to participate (e g , one man, one vote) Perhaps the primary reason that this form of ex ante equality has been stressed in the literature and institutions of democracy is that it is considered a basic sort of right, or value judgment, which goes hand in hand with political democracy (Vlastos 1962, Tobin 1970) [4] A second important reason for one man, one vote could be that there are only minor disincentive effects associated with guaranteeing this form of political equality.[5]

Given the relative lack of disincentive effects associated with mechanisms for reallocating voting rights, it is interesting that the discussion of political equality has never seriously considered establishing an ex post equality of the distribution of political income.[6] In markets for private goods, each voter-consumer allocates his income so that the expected marginal utility from each dollar vote on a good is equal. Planned consumption at equilibrium prices is identical with actual consumption. In the absence of mistakes due to lack of sufficient information, this assures equality between expected and actual utility These conditions are not likely to hold where public goods are allocated via one-man, one-vote majority rule The utilities a voter expects from each vote may vary widely And there can exist a large divergence between what he expects (hopes?) to receive in public goods as indicated by his voting decisions and what he actually receives after all votes are cast and counted

---

[3] For Lerner's proposal and some discussion of it, see Lerner (1935, 1970), Breit and Culbertson (1970, 1972), Morrison and Higgins (1971), McCain (1972), and McManus, Walton, and Coffman (1972)

[4] Tobin (1970, p 269) also argues that "obviously an equalitarian distribution of the vote can be enforced without any loss of efficiency The aggregate supply of votes is intrinsically inelastic " This is an overstatement of the issue, since although voting can be enforced or subsidized and in the extreme all would vote, this does not guarantee that votes are of any given level of informational quality We have discussed elsewhere some ways out of this dilemma See Mueller, Tollison, and Willett (1971, 1972*a*, in press)

[5] For example, an alternative distribution of the right to participate in the political process might be to give votes proportional to years of formal education, i e , some sort of scheme of weighted voting rights (Vlastos 1962) Counteracting the need for this sort of system, however, are such factors as the effect of education on one's perception of his "obligation" (elasticity) to participate in politics which tend to lead to very few observed disincentive effects for educated voters under a system of full, ex ante political equality with everyone voting and weighted equally Thus, on practical grounds political equality of this sort could be defended For empirical evidence on this point, see Milbrath (1965) For a discussion of the economic theory of voting behavior, see Tollison and Willett (in press)

These two characteristics of voting have been typically described as giving rise to problems of "minority rights" or the "tyranny of the majority " It is important to note, however, that these characteristics create *two separate problems*, and thus require separate remedies Let us illustrate their differences with an example

Suppose a community plans to build a town hall A referendum is held on whether the building should be of Gothic or Greek architecture Fifty-one percent of the voters have a slight preference for the Gothic style, the other 49 percent are strongly opposed On a straight majority vote, the Gothics win even though their utility gain is far outweighed by the minority's loss We can describe this as a situation violating the principle of ex ante equality Each citizen's vote does not represent an equal potential utility change

Consider next the case where each citizen's vote does represent an equal potential utility gain Then a simple majority rule will maximize the net utility gain for the community on this decision Suppose, however, the town grows and new public buildings are erected Each time the town votes on the architecture of the next building, and each time the pro-Gothics win Every decision will be *efficient* in the sense that the utility gains of the winners exceed the losses of the minority, but one might well regard the long-run outcome as *inequitable*, as all the utility gains accrue to the pro-Gothics (ignoring the utility losses both groups might experience from mixing Gothic and Greek revival buildings) Thus, even where there is full ex ante political equality, there is still the possibility of ex post inequality of the utility income derived from the political sector A person consistently in the minority may receive ex post a far smaller share of the gains generated by public collective action, even where there is ex ante equality over all voters[7] Indeed, we would expect considerable ex post inequality of political outcomes to be the general case, rather than the exception.

---

[6] Braybrooke (1968) is a partial exception In his proposed tests for whether actual governmental systems conform closely enough in practice to the theoretical norms of democratic theory to be approximately labeled democracies, Braybrooke suggests the criteria that no voter be in the minority more than 90 percent of the time and that not more than 10 percent be in the minority more than two-thirds of the time (p. 171).

Can political systems be devised to achieve both ex ante and ex post equality of political income? In the following section we consider mechanisms which seek to achieve this objective

II. Solutions

The problem created by differing intensities of preference among voters can be solved by introducing point voting Since point voting has been discussed in detail elsewhere, we shall only briefly outline it here (Musgrave 1959, pp 130–31, Mueller, Tollison, and Willett 1971, in press) To begin with, voters and issues must be grouped into polities such that in any given polity each voter has an "equal stake" (in terms of his potential aggregate utility change) in the *set* of issues decided by the polity[8] That is, governmental units are broken up into, say, neighborhood, city, regional, and national jurisdictions, so that the integral of voters' utility distributions over the issue set to be decided by a given polity is approximately equal Each member of a polity is given an equal number of vote points to be allocated over the set of issues to be decided by the polity If the voters honestly allocate the vote points in proportion to their relative intensities of preference on all issues (i e , they do not engage in strategic behavior), the point allocations will reflect each voter's marginal rates of substitution between issues[9] If the equal-stake criterion is met, the total expected utility of each voter in a polity will be equal, and the outcomes obtained from aggregating the point-vote allocations will maximize the utility gains of the community from voting The equal-stake criterion produces ex ante political equality for each voter over the set of issues, while point voting allows each voter to achieve ex ante equality on every issue by revealing his relative intensities of preferences over the issue set Together they form a system for achieving ex ante political equality for each voter on every issue

There are two contexts in which one can conceive of equalizing the distribution of political income ex post First, one could adjust each individual's vote stock at the beginning of a voting period, depending upon his success over *all* previous periods Suppose that in the first round an individual won on issues totaling 90 percent of his vote points, while the average voter won only 70 percent. If the initial stock were 100 votes, the individual would be given 80 votes at the start of period 2. If he again won on 90 percent, he would win on 72 vote points, and his tax for the next period would be two votes He would thus have 78 votes for period 3, and as long as he continued to win on 90 percent of his vote point allocations, he would be winning on issues totaling 70 vote points, the average for the community If, on some round, either his tastes or the community's changed so that he only won on half of his allocations, his total successful allocations would fall to 39 In the next round his cumulative overall tax would be 20 + 2 − 31 = −9, and he would receive 109 votes If he continued to win on only 50 percent of his vote-point allocations, his vote holdings would continue to rise until he began receiving 140 vote points, assuming the average voter was still winning on 70 vote points [10]

---

[7] The point here is similar to Arrow's (1971) discussion of input and output equalization for public expenditures Also, evidence from simulation studies of the gains from trade from vote trading indicate that the quantitative importance of the problem of political income distribution may be significant (Mueller, Philpotts, and Vanek 1972)

[8] For analytical discussion of the equal-stake criterion, see Mueller (1971)

[9] Mueller (1973) discusses the conditions under which point voting proceeds honestly

This procedure is a compromise between the ex ante equalitarian condition embodied in the equal-stake criterion, and a full ex post equality of political income as described below The stability of voter preferences, and therefore of voter winning percentages, is relied upon to adjust their vote holdings so that each voter's expected successful vote allocations, based on past outcomes, are the same for the coming round (If the voters' winning percentages are not stable over time, there is no long-run tyranny of the majority problem, and hence no need to adjust individual vote stocks) In the above example, the voter would continue to get 78 votes as long as his winning percentage stayed at 90, because he would then continue to win on issues totaling 70 vote points, the community's average No adjustments would ever be made, however, for the 22 "extra" vote points he won in the first two rounds He is always "ahead of the game" by that amount Alternative tax-subsidy schemes could be designed to achieve equal *cumulative* ex post political income, but these would have the serious disadvantage of requiring a constant potential total utility gain for each voter from period to period [11]

---

[10] One might think at first glance that the tax would not have to be the integral of all past deviations from the median, but just the divergence since last period But such a system does not converge to the median, since the voter is only taxed the incremental divergence of winning or losing votes from the median in the immediately preceding period In the above example the individual's tax at the end of two periods would only be two votes, allowing him a total of 98 votes as opposed to 78 votes where the integral of differences over all past votes is taken

[11] This problem could possibly be handled by introducing equal increments of votes say each voter in the polity gets 10 additional votes in the period in which a new public-good issue (of importance roughtly equal to 10 percent of the normal total) is introduced Then the vote-tax system can continue, and "new money" in the form of the extra votes will keep "purchasing power" proportionate to total income Such a solution would follow, of course, if voting periods were time dependent An alternative way to meet this problem is to make voting periods not time dependent, but dependent on the total value of issues, e.g , vote every *x* issues worth of political income However, the problem of holding the total utility gain constant for each voter from period to period would involve

No such restrictive assumption is required for the lagged vote-tax scheme, since its goal is merely to equate each voter's expected winning number of vote points in the coming period By basing these expectations on each voter's past history of successes, and taking into account the stability of these winning percentages, individual vote stocks could be adjusted to bring ex ante and ex post political income into equality over time.[12] The voter who has done badly in the past can only write these losses off, and console himself with the knowledge that he may expect a more equal share of public-good pie in the future, whatever the size of that pie might be [13]

Alternatively, one might decide to equalize the realized utility gains over the set of issues a polity decides at any point in time, by a system of taxes and subsidies on voting rights and a sequential, nonbinding series of votes on issues Such a system could achieve *full* ex post equality of political income Initial votes would be tentative, subject to the vote-income distribution that resulted from voting As an example, suppose that after the initial vote the set of outcomes would be such that each voter would have on average decisions corresponding to 70 percent of their vote-point allocations satisfied If a particular voter won on a set of issues corresponding to 90 percent of his vote-point allocations after the first round, then his total vote holdings would be reduced for the second round of voting Thus, vote stocks would be adjusted by taxing abnormally high winners (say a tax of 20 votes on the 90 percent winners) and subsidizing losers, after which the vote points would be cast again Over a series of recontracts with adjusted stocks of vote points, the gains from winning and losing would be equalized To avoid excessive recontracting

---

gains and losses analogous to the windfall gains and losses that accrued in the first few periods to above-average winners and losers That is, changes in the size of the pie would create additional gains and losses until the lagged vote-tax system equated expected voters' winnings over a longer run To handle either of these problems neatly would probably make the vote-tax system unduly complicated and also raise additional conceptual problems

[12] Of course, where there are strong differences in preferences on an issue (i e , an intense majority facing an intense minority), the system of political taxation may never succeed in equalizing income For that matter, the equal-stake criterion may not be satisfied Such issues should rationally be decided under conditions of equal ignorance about the future in a constitutional period, and if agreement cannot be reached there, then separate states should be set up See, in particular, Mueller (1971, 1973) and Mueller et al (1971, 1972*a*, 1972*b*, in press)

[13] Problems exist in applying the lagged vote-tax system to a representative body If the legislators represent geographic units, a newly elected member could simply inherit the stock of vote points of his predecessor. This would introduce a certain amount of instability if newly elected representatives had significantly different tastes than their predecessors Redistricting would also cause problems, as would a proportional representation scheme as discussed in Mueller et al (1971, in press) These problems vanish if immediate recontracting is employed

under this scheme, one might be satisfied with bringing everyone within five or 10 points of the average utility gain.[14]

Thus, as is illustrated by these examples, operational systems of vote taxation that attain greater equality of political income can be devised and applied in representative democracies The importance of such a mechanism in a well-working model of representative democracy lies in its ability to reduce or eliminate the potential for a tyranny-of-the-majority problem to persist over a long period of time.

III. Summary

This paper has discussed possible mechanisms for achieving ex ante and ex post equality of political income While any of the structural reforms could be implemented separately, it makes the most sense to combine them Indeed, the "equal-stake" criterion is to some extent a necessary condition for the implementation of a redistributive vote tax A voter's winning percentage on an issue set cannot be regarded as an index of his political income unless each vote has an equal value ex ante to all voters [15]

The idea of the state arising from a "social contract" formed by all citizens to achieve collective ends can imply the concepts of both ex ante and ex post political income equality Of the two, only ex ante equality has received attention in the literature Yet, to those who embrace the equalitarian ethic, the ex post equality of political income should be of even greater importance than ex ante equality of voting rights This paper argues that the attainment of ex post equality of the political income distribution is conceptually plausible and quantitatively important

Further, a policy of vote taxation does not appear to be subject to as serious disincentive effects as taxation to obtain complete equality of ex post private economic income This follows by analogy to elasticity conditions where voting takes very little time as opposed to the length of the work week, for example, and attempts to equalize realized private income

There do not appear to be any functional equivalents under present democratic systems to achieve ex post equality of political income [16]

---

[14] The recontracting procedure can be simplified by the use of a computer The voter need reveal only once the percentage of his vote stock he assigns to each issue (i e , his relative intensities over the issue set) The computer can then adjust stocks to bring each voter's realized gains into equality

[15] It should also be noted that the application of the vote-tax system relaxes the problem of assuming that identical tastes obtain for the distribution of public goods in models of simultaneous determination of optimal distribution and optimal allocation with both private and public goods See, e.g , Winch (1969, 1970) and Roberts (1970).

[16] Approximations in the private sphere can be found in the way competition is set up in some sports Perhaps the handicapping system in golf is the best example corresponding to the vote tax However, such institutions as the player draft in professional sports show much the same purpose as the vote tax in the sense that successful teams are penalized

Indeed, taking the United States as an example of a two-party, geographic-based democratic system, one could argue that there are perverse mechanisms, such as monopolistic systems of committee power and the like, which tend to lead to greater inequality of realized utility for many voters Many voters from lower income-education classes do not vote under the present system, which, although it does not mean that these voters do not receive political incomes, does mean that the equalization of political income would be a difficult task without also implementing a thorough reform of public choice mechanisms to insure efficient revelation of *all* voter preferences in polities defined on the equal-stake criterion

References

Arrow, K J "Equality in Public Expenditure " *Q.J E* 85, no 3 (August 1971) 409–15

Braybrooke, D *Three Tests for Democracy* New York Random House, 1968

Breit, W., and Culbertson, W P , Jr "Distributional Equality and Aggregate Utility Comment " *A E R* 60, no 2 (June 1970) · 435–41.

——— "Distributional Equality and Aggregate Utility Reply " *A E R* 62, no 3 (June 1972) 501–2

Fair, R C "The Optimal Distribution of Income " *Q J E.* 85, no. 4 (November 1971) 551–79

Kolm, S C "Justice et équité " Mimeographed Paris CEPREMAP, 1971

Lerner, A P *The Economics of Control* New York Macmillan, 1935

——— "Distributional Equality and Aggregate Utility Reply " *A E R* 60, no 2 (June 1970) 442–43

McCain, R A "Distributional Equality and Aggregate Utility Further Comment " *A E R* 62, no 2 (June 1972) 497–500

McManus, M , Walton, G M , and Coffman, R B "Distributional Equality and Aggregate Utility Further Comment " *A E R* 62, no 2 (June 1972) 489–96.

Milbrath, J *Political Participation* Chicago Rand-McNally, 1965

Morrison, C C , and Higgins, R S "On Distributional Equality " *Southern Econ J* 38, no 4 (October 1971) 260–62

Mueller, D C "Fiscal Federalism and Constitutional Democracy " *Public Policy* 19 (Fall 1971) 567–94.

——— "Constitutional Democracy and Social Welfare " *Q J E.* 87, no. 1 (February 1973) 60–80

Mueller, D C , Philpotts, G , and Vanek, J "The Social Gains from Exchanging Votes " *Public Choice* 13, no 2 (Fall 1972) 55–79

Mueller, D C , Tollison, R D , and Willett, T D "Solving the Intensity Problem in Representative Democracy " Mimeographed Cornell Univ , 1971

---

after the season with lower draft picks in an effort presumably to make next season's competition more equal ex ante The ability of teams to trade personnel and their lack of ability to forecast perfectly future players' performance hinders the attainment of full ex ante equality (in the sense that each team becomes equally likely to win the title) over time in such sports (thus the Green Bay Packers, Boston Celtics, and so forth) However, there do not appear to be any equivalent reasons to think that the vote tax would operationally generate long winning or losing streaks Indeed, it is to avoid such winning streaks in the political process that one would advocate a vote tax

———. "Representative Democracy via Random Selection." *Public Choice* 12, no. 1 (Spring 1972): 59–68. (*a*)

———. "The Utilitarian Contract A Generalization of Rawls' Theory of Justice." Mimeographed. Cornell Univ., 1972. (*b*)

——— "A Normative Theory of Representative Democracy " In *Quantitative Methods of Representation* New York: New York Acad. Sci., in press.

Musgrave, R *The Theory of Public Finance.* New York: McGraw-Hill, 1959.

Pauly, M. V , and Willett, T. D. "Two Concepts of Equity and Their Implications for Public Policy." *Soc Sci. Q.* 53 (June 1972). 8–19.

Roberts, D. J "Pareto, Public Goods, and Politics." *Canadian J. Econ* , no. 3 (November 1970), pp 615–16

Thurow, L. "The Income Distribution as a Pure Public Good " *Q.J.E.* 85, no. 2 (May 1971) 327–36.

Tobin, J. "On Limiting the Domain of Inequality " *J Law and Econ.* 13 (October 1970) 263–79

Tollison, R D , and Willett, T D "Some Simple Economics of Voting and Non-Voting." *Public Choice* (in press)

Vlastos, G "Justice and Equality " In *Social Justice*, edited by R. B Brandt New York Prentice-Hall, 1962

Winch, D. M "Pareto, Public Goods and Politics " *Canadian J Econ.*, no 2 (November 1969), pp 492–508

——— "Pareto, Public Goods, and Politics Reply " *Canadian J Econ* , no 3 (November 1970), pp 616–17

Some Principles of Inflationary Finance

Donald A. Nichols

University of Wisconsin

It is well known that a government can use inflation instead of taxation as a means of financing its expenditures.[1] The amount of inflationary finance possible has been shown to depend on the demand for money Previous work, however, has ignored the fact that a government can shift the location and elasticity of that demand curve through the use of various controls and regulations, and thereby change the level of inflationary tax receipts that is associated with a given level of inflation I will argue here that exchange controls, interest rate restrictions, and debt management policies all affect the demand for money and the possibilities for inflationary finance Since a well-established model of inflationary finance exists in the literature cited, I will confine myself to an analysis of the links between the policies in question and the demand for money, and will point out that both an increase in the demand for money and a reduction in its elasticity with respect to the nominal interest rate will increase inflationary tax possibilities in the traditional model [2]

I

The demand for money obviously depends on whether good substitutes exist With good substitutes easily available, it is likely that the demand for money will be less than otherwise and more elastic with respect to the rate of interest [3] Thus we are led to our first principle, which I believe is the rationale for many of the capital market impediments we observe in many countries today

The inflation that results from a given government deficit will be lower than otherwise if the availability of close substitutes for government money is curtailed

---

The author gratefully acknowledges the financial support of the National Science Foundation and the Graduate School of the University of Wisconsin The research was completed while the author was a visitor at the London School of Economics He wishes to thank Harry Johnson and Marcus Miller for their useful comments on an earlier draft of this work

[1] See, for example, Bailey (1956), Mundell (1965), Marty (1967), Sidrauski (1967), and Friedman (1971)

[2] The model is spelled out in detail in a previous version of the paper available from the author

[3] See Tobin and Brainard (1963) for a discussion of monetary policy where good substitutes for money exist

One good substitute for domestic money is foreign exchange. If we assume the existence of flexible exchange rates, long-run equilibrium requires that the exchange rate change at a rate equal to the difference between the rates of inflation of the domestic and the reserve currencies. One way for a money holder to avoid the inflationary tax in this situation is to hold all money balances in foreign currency and to use domestic currency only when it is absolutely necessary. Thus with foreign currency a good substitute for domestic currency, an immediate corollary of the above principle is·

The rate of inflation will be lower if the holding of foreign currency is restricted or prohibited

The exchange controls needed to police this prohibition are not for the purpose of restricting imports; free trade can still be allowed Rather, the purpose is to prevent the decline in the demand for domestic money—and therefore of inflationary tax revenue—that would occur in the presence of a free currency market It is the holding of foreign assets that must be curtailed, not the consumption of foreign goods

Private assets can also compete with public ones in individual portfolios. If the private assets bear a market-determined interest rate, and if that monetary interest rate increases with the long-run rate of inflation, then we would expect large portfolio shifts from currency into the competing private assets as the rate of inflation increases In this case again, a higher rate of inflation will result from a given government deficit than would occur if these private assets were restricted. One way to limit the appeal of these competing assets is by restricting the rates of interest they may legally bear. Once the legal ceiling is reached, further increases in the rate of inflation will not increase the relative attractiveness of private over public assets.[4] Of course, an alternative to reducing the attractiveness of private liabilities is to require that their issuers hold public currency. Thus a second corollary emerges.

Restricting the interest rates legally payable on private assets will decrease the rate of inflation that results from a given government deficit.

Should these restrictions be successful in curtailing the availability of good substitutes, we would expect the demand for money to become more inelastic with respect to the rate of inflation than it would otherwise be. One instance that seems to support this prediction is that of Brazil in the last two decades Kafka (1967) reports that during this period, the Brazilian money supply remained a reasonably constant fraction of annual gross domestic product (about 20 percent) despite very large changes in the rate of inflation. Capital market restrictions of the kind I discuss above were in use at that time and may explain the inability of the Brazilian money holder to escape the extremely large taxes which were levied on his cash balances.

---

[4] See Tobin (1970) for a discussion of monetary policy with interest ceilings on private assets.

Thus governments need not take the demand for money as fixed when determining their inflationary policies They can affect it through legal restrictions. Thus the estimates provided by Friedman (1971) of the rates of inflation that maximize inflationary finance must be interpreted in this light. They should be thought of as rough lower bounds of the true estimates, since they are applicable only when legal restrictions of the kind just mentioned are unavailable. The true revenue-maximizing rates cannot be calculated in the absence of knowledge of the degree of substitutability of other assets for money

II

Consider next a government which issues two liabilities, one interest-bearing and one non-interest-bearing If the non-interest-bearing liability is currency, it is interesting to ask what the characteristics of the interest-bearing asset should be if all possibilities for inflationary finance are to be fully exploited The answer to that question is seen most easily in the framework outlined by Tobin in his essay on debt management [5] We can state it in the form of a principle

A government can lower the rate of inflation caused by its deficit if its interest-bearing debt is made to be a poor substitute for its currency and a good substitute for other assets in private portfolios

What characteristics of debt make it a better substitute for private assets than for money? The answer depends on what the private assets are like, of course, but since we have already shown that an inflationary policy requires that private financial assets be restricted, the remaining assets are likely to be real—either consumption or investment goods, equities which are claims to goods, or land We wish to give the interest-bearing debt some characteristics which allow it to satisfy the same portfolio objectives as goods, and to avoid those characteristics which allow it to satisfy the same objectives as money.

The debt might be made illiquid, perhaps nonnegotiable, and of long maturity to make it a bad substitute for money It might be given an intertemporal pattern of payouts which dominates that of goods for certain purposes—a retirement annuity, for example, might be preferred to real capital goods by many asset holders Many different kinds of debt could be issued, of course, to satisfy the different objectives of savers. Interest-bearing debt should be made very useful to wealth holders yet a very poor

---

[5] See Tobin (1963). This paper owes a great deal to the work of Tobin, since he, more than anyone else, has developed the policy implications of having assets be imperfect substitutes for each other in individual portfolios.

way of storing short-term purchasing power if the purpose of the debt is to lower the rate of inflation that results from a given fiscal deficit.

One result established in the previous literature is that deficits financed by money cause inflation only to the extent that the rate of growth of the money supply exceeds the rate of growth of the real economy. At any long-run equilibrium, all forms of government debt can be expanded at the rate of growth of output without any changes in relative asset prices. This implies that even an interest-bearing debt will yield steady-state revenue if the rate of growth of the economy exceeds the rate of interest the debt bears The interest rate on each form of debt could be chosen to maximize the revenue derived from growth plus inflation for the total debt. Note that a purchasing-power guarantee could make the debt preferable to some private assets, yet if the real interest rate is kept below the growth rate, the debt could still yield steady-state revenue

III

Foreign aid is sometimes mentioned as a useful device for slowing an inflation It should be obvious, however, that only permanent foreign aid that grows at the rate of growth of the economy can stop a permanent inflation Ultimately, a country must tax itself if it wishes to finance its government expenditure in a noninflationary fashion and if it is unable to secure perpetual foreign aid of the appropriate size

Foreign aid cannot stop an inflation permanently unless the aid is given forever.

A puppet regime might be supported by perpetual foreign aid. Puppet regimes are often too unpopular to raise tax revenue domestically, they can stay in power only if they make a sufficient level of government expenditure, and these must be financed from abroad

Loans that must be repaid with the lender's currency will have a long-run beneficial effect on the inflation only if the interest rate charged net of the lender's rate of inflation is less than the rate of growth of the domestic economy Otherwise, steady-state repayments will ultimately exceed receipts of new loans Repayments obviously have the reverse effect on the inflation to that of aid Loans specified in the borrower's currency can yield steady-state revenue if the rate of growth exceeds the stated interest rate minus the rate of domestic inflation.

IV

I have ignored so far the question of private capital accumulation. I have been concerned only with a government that wishes to find some means of financing its own given level of projects. Briefly, I will now sketch in some of the considerations that should influence the government that is interested in furthering certain private forms of investment as well. Those underdeveloped countries, for example, for whom much of this analysis is most relevant wish that the public investment projects they undertake replace private consumption rather than private investment. In this case, it is not sufficient to worry about only minimizing the rate of inflation that is caused by the government expenditure, since other goals are important as well.

I will concern myself here only with private investment Mundell (1965) and Marty (1967) have analyzed elsewhere the case in which the government chooses its financing as a function of the investment opportunities it faces Clearly, in that case it must also consider the extra future revenue to be yielded by a project when determining the profitability of further investment and finance Then it is the rate of return on public projects which becomes the maximum interest rate payable on government debt, not the rate of growth [6]

In the analysis above, we were concerned with the degrees of substitutability in private portfolios between goods in general and different forms of government debt If private investment in producer durables is also a government goal, then the characteristics of the various forms of debt it issues should be chosen so as to make them substitute for consumption goods, not investment goods In this case, debt should be tailored so as to fit well in private portfolios that otherwise would hold no plant and equipment or other assets that traditionally finance plant and equipment. These portfolios are likely to be the smaller ones

To illustrate, consider the different effects on investment of two alternative forms of debt the government might issue (*a*) a small-denomination retirement annuity bearing a very low (perhaps slightly negative) real rate of interest, or (*b*) a large-denomination bond bearing a monetary interest rate above the legally allowed rate on private debt Both assets *a* and *b* might be better substitutes for goods than money, and therefore either of them could be used to finance the government's projects But asset *a* is likely to replace the purchase of gold, transistor radios, or motor bikes, while asset *b* might well replace bank deposits, private loans, or other debts of productive organizations Clearly, asset *a* is the one the government should issue if it is concerned with the level of private plant and equipment investment Debt management policy should make the government debt substitute as little as possible for funds that would be spent on other socially desirable projects

It should be noted that the government could also lend money to any worthwhile organization it feared had been harmed by its financial policies. Such loans have the reverse effect of government borrowing, of course, but as part of a whole program, they may be counterinflationary. That is, if the willingness to make such loans enables the government to issue a form of debt it would otherwise avoid, the inflation may be countered by the lending policy With interest rate restrictions, private borrowers profit from an inflation too, of course, so the willingness to lend to certain productive organizations is one way of providing them with the incentive of a subsidy. Alternatively, should certain forms of debt be used traditionally to finance undesirable projects, it is these that the government should restrict and replace by its own debt

---

[6] Since the proceeds of debt finance in my analysis are to be used to reduce taxes, it is the rate of growth rather than the rate of return which limits the interest rate payable on government debt if it is to yield revenue in the steady state See Diamond (1965) and Cass and Yaari (1966) for proofs of this in different contexts

One of the common victims of inflation is the savings of the poor. Indeed one could interpret the analysis above as a set of devices whose purpose is to marshall the savings of the poor for national development purposes But it may be true that the poor's elasticity of saving with respect to the real rate of return is so great that saving goes to zero once a large inflation is permanently established, or, more accurately, that all saving is then done in the form of durable consumer goods If this is the case, then the inflation has caused to be wasted a large potential source of saving. What is needed in this case is a new form of government debt which will restore the profitability of saving to small savers yet which will not force the government to abandon the profits of inflationary finance

Examples that might be tried abound and are limited only by the imagination of the government A ticket of admission to a retirement home, for example, is a real debt that might be sold in small parts at a rate that gives quite a profit to the government, yet which appeals to private asset holders Claims to food during retirement, a vacation, or some durable good to be delivered in 5 years time all are assets that are guaranteed against inflation, yet which can be sold at high enough prices to make them profitable for the government Debt, in short, can be made to substitute for many things besides money and productive investment

V

In conclusion, it is useful to place in perspective the kinds of controls I have discussed, but not necessarily recommended Economists have long been in favor of untrammeled markets and have usually been critics of the controls I describe. Yet we should judge these controls with respect to the practical alternatives available, not according to some theoretical ideal. All practical forms of taxes have welfare and collection costs. All have loopholes which must be closed at some administrative expense. Viewing inflation as a tax on the holders of government money, we can then view the ready availability of substitutes as a loophole for avoiding the inflationary tax. Closing the loophole takes some administrative effort and requires that we prevent certain transactions from taking place. But since this is true of all taxes, it becomes an empirical question whether the restrictions are more costly for inflation than they are for other forms of taxes.

Previous estimates of these costs indicate that the question is at least an open one A true calculation of the costs could best be done in a larger model with a social welfare function. Only more limited partial analyses have been carried out The cost of inflation as shown by Bailey (1956) derives from the fact that with inflation, asset holders hold less money than they otherwise would and therefore receive less of the benefits that are derived from money holding Sidrauski (1967) has shown that this can lead to more or less steady-state consumption of other goods Marty (1967) argues correctly that there is no a priori way to tell whether these costs exceed the efficiency and collection costs of other practically useful forms of taxation He compares estimates which show the efficiency costs to be about the same for inflation as for other forms of taxes. The policies I have described have obvious efficiency costs, but equally obviously increase the revenue available to the government Unless their costs can be shown to be unusually large, they merit consideration as practical possibilities for financing government expenditure.

Finally, this analysis warns us that where capital market restrictions already exist, premature removal of them as part of a counterinflationary policy may actually accelerate the inflation if other steps are not taken to offset their impact on the demand for money. The elimination of capital market restrictions may be one of the goals of a counter-inflationary policy, but it is not one of the tools to accomplish that end.

VI

I have tried to outline some of the principles of inflationary finance The purpose was to show that many different rates of inflation may result from a given government deficit The roles of exchange controls, bank regulation, and capital market restrictions were discussed as part of a broad policy of inflationary finance Debt management policies were described as well as the interaction between the policy of inflation and other goals New forms of debt were suggested as being likely to provide finance to the government without hindering other objectives

I have considered only balanced growth situations In many instances, however, the adjustment from one steady state to another provides an opportunity for one-shot financing that may be of substantial size A social security scheme, for example, that pays an interest rate equal to the real rate of growth collects no net revenue in a steady-state equilibrium But the introduction of such a scheme can provide a massive inflow of revenue in the short run, and such a device should be considered as a useful way of overcoming the problems of adjusting from one steady-state rate of inflation to another—in this case—lower one. Other permanent policies will have to be invoked simultaneously, of course, if the new lower rate of inflation is to be permanently maintained.

References

Bailey, Martin J "The Welfare Cost of Inflationary Finance." *J P E.* 69 (April 1956) 93–110

Cass, David, and Yaari, Menahem F. "A Re-Examination of the Pure Consumption Loans Model." *J.P.E.* 74, no 4 (August 1966) · 353–67

Diamond, Peter A. "National Debt in a Neoclassical Growth Model." *A.E.R.* 55, no 5 (December 1965): 1126–50

Friedman, Milton "Government Revenue from Inflation" *J P E* 79 (July 1971) 846–56

Kafka, Alexander "The Brazilian Stabilization Program, 1964–66." *J P E.* 75, no 4, suppl (August 1967) 596–631

Marty, Alvin L "Growth and the Welfare Cost of Inflationary Finance." *J P E.* 75 (February 1967) 71–76

Mundell, Robert A "Growth, Stability and Inflationary Finance." *J P E* 73 (April 1965) 97–109.

Sidrauski, Miguel "Inflation and Economic Growth." *J P.E* 75 (December 1967) 796–810.

Tobin, James "An Essay on Principles of Debt Management" In *Fiscal and Debt Management Policies* Commission on Money and Credit Supporting Papers Englewood Cliffs, N J Prentice-Hall, 1963

——— "Deposit Interest Ceilings as a Monetary Control" *J Money, Credit and Banking* 2, no 1 (February 1970) 4–14

Tobin, James, and Brainard, William C. "Financial Intermediaries and the Effectiveness of Monetary Controls." *A E.R* 53, no. 3 (May 1963). 383–400.

A Bias in Current Measures of Economic Growth

Milton Friedman

University of Chicago

It is common practice to measure the growth in economic welfare in a country over any considerable period by the rate of growth in real per capita income. However, this measure can be seriously biased for a period during which the country experienced substantial immigration or emigration.

Consider the United States from, say, 1870 to 1914. During that period, real per capita income as measured by Simon Kuznets grew at the annual rate of about 2 percent. However, there was heavy immigration, reaching a peak rate of over 1¼ million persons a year in 1907. By 1914, roughly one-third of the total population of the United States was foreign born or of foreign or mixed parentage, and one-fifth of the population over 14 years old was foreign born. Their counterparts in 1870—for, of course, we want to measure the growth in economic welfare for a "representative" population, not for a collection of identical and aging individuals—lived or worked in other countries, not in the United States. We have every reason to believe that their 1870 counterparts had substantially lower incomes than the then population of the United States—presumably that is why the United States experienced heavy immigration, not emigration.

The foreign-born in 1914 probably also had lower incomes than the rest of the U.S. population, but again the fact of continued large-scale immigration, let alone a wide variety of historical evidence, suggests that their incomes in the United States were substantially above the incomes that they could have received in the countries from which they came. In any event, the incomes of the 1914 foreign-born are included in the aggregate income underlying the 1914 per capita income estimate, while the incomes of their counterparts in 1870 are completely excluded. The result is to bias the estimated rate of growth downward. Excluding the foreign-born in 1914 as well as their counterparts in 1870 would raise the 1914 per capita income figure and leave unchanged the 1870 figure. Including the foreign-born in 1914 and also their counterparts in 1870 would leave unchanged the 1914 figure while lowering the 1870 figure. Either way, the rate of growth would be raised.

For simplicity, I have stated the problem simply in terms of foreign-born, but clearly a change in the population of first-generation, second-generation, etc., residents would also affect the results.

One of the great economic achievements of the United States in the period from 1870 to 1914 was the absorption of millions of residents who came to the United States with little but their bare hands, were able to make a better life for themselves than in their countries of origin, and to lay the foundation for a still better life for their children Yet not only is this achievement not recorded positively in the common measure of economic growth, it actually enters as a negative factor, reducing the measured rate of growth.

As a purely arithmetical matter, it would, for example, be possible for the average income of all residents of a country to remain constant, even though the income of every person separately was rising, provided that the immigrants entered at an income lower than the average, though higher than their base income in the country of origin, and that either the income at which they entered was falling relative to the average or the percentage of foreign-born was rising sufficiently rapidly.

It is almost surely impossible to correct the bias by including the counterparts of the 1914 foreign-born in the initial income estimates In principle, however, the bias could be corrected with U S data alone by using initial and terminal average incomes for a population standardized not only with respect to distribution by sex and age, but also by length of residence in the country of the individual himself or his forebears I do not know whether data exist that would permit such standardized averages to be computed even roughly

What seems clear is that such standardized averages would show a more rapid rate of rise in real per capita income in the United States over the period 1870 to 1914 than the current crude averages And it is possible that the difference would be substantial

A similar problem arises today in the income estimates for countries using or providing much foreign labor How should the more than one-half million Yugoslav workers now employed in Germany be allowed for in the income estimates for Germany? for Yugoslavia? How are they? I do not know the answers, but any serious student of rates of growth in economic welfare in Germany and Yugoslavia in the postwar period must face these questions

Miscellany

Human Migration: A Survey of Preclassical Literature

The study of human migratory behavior is not recent in the studies of social analysts In fact, a thorough review of the literature reveals information concerning this phenomenon even in the Old Testament The plight of the ancient Hebrews is a case in point. Prior to their residence in Egypt, the Jewish religion found its mystical roots in the area of current Israel However, due to increasing economic hardship resulting from years of drought (7 to be exact, witness the case study of Joseph of the "coat of many colors" [Gen , chaps 37–46]) Jacob and his son and relations migrated to Egypt in search of an easier and more prosperous life (Gen 46 1) This migration was not made to a totally unknown environment, because a member of Jacob's family, Joseph (mentioned above), was already an established member of the Egyptian community, holding a position comparable to secretary of agriculture Furthermore, an omniscient information source also recommended the migration (Gen 46:3) In fact, the migration did improve welfare, although most notably the long-run welfare of the Egyptians increased due to enslavement of the Jews.

> And the Lord said, I have surely seen. the affliction of my people which *are* in Egypt, and have heard their cry by reason of their taskmasters, for I know their sorrows . Now therefore, behold, the cry of the children of Israel is come unto me. and I have also seen the oppression wherewith the Egyptians oppress them. [Exod. 3.7, 9]

Thus, in this case the long-term consequences of migration were economic and social subjugation.

Several generations later, under the leadership of a man named Moses (who supposedly talked to God), the Jews migrated from Egypt once again to obtain a substantial increase in their material and spiritual welfare The destination was reputed to be highly "promising"—"a land flowing with milk and honey" (Exod. 3·8). In this case, the probability of obtaining the expected income stream in the destination was certainly in that the Source of their information was once again unquestionably reliable. (But, come to think of it, He suggested that they go to Egypt in the first place.) Well, in any event, even if the information was not of perfect reliability, the subjective probability estimate for the relative income stream differential between the origin, Egypt, and the "Promised Land" was no doubt sufficient to induce migration. A problem exists with regard to relating the migration decision to the distance between the origin and destination in that the route they traveled was highly circumscribed and the time distance was not known prior to migration. No doubt had most of the Jews known that the time distance was 40 years (Deut 1·3), this would have compensated for the increased economic welfare in the "Promised Land" and perhaps migration would not have occurred.

In addition to the standard neoclassical welfare maximization theory of migration, the Old Testament provides two other explanations for migration The first is that of ostracism One sociological explanation for migration is that the migrant has been ostracized by the dominant group in the population of a given geographic area The reason for the ostracism may vary but most commonly involves the violation of the dominant group norms The Garden of Eden is an illustrative example (Gen., chaps 1–3) It is clear that in this case God represents the dominant "group " It is also clear that Adam and Eve violated one of the primary norms of behavior in the Garden of Eden in that they ate of the fruit of the Tree of Knowledge (Gen 3.6) Thus, grounds for ostracism existed

However, the mere observation that they violated the norms of "the first society" is insufficient, we must ascertain the reasons for the violation A more detailed analysis of the text reveals that shortly prior to the norm violation, the migrants, Adam and Eve, interacted with a member of a deviant minority subculture in the garden, the snake (Gen 3:1–5). This member of a deviant subculture disregarded the norms of God and sought to get Adam and Eve to do the same However, mere association with the snake would not have been sufficient if Adam and Eve had not desired to rise in the social structure by increasing their level of knowledge. This implies that their reference group was not the same as their face-to-face group in the garden Rather, it might be suggested that they wished to become members of the face-to-face group of which God was a member (Gen 3 5) It is clear that God had great knowledge, and perhaps Adam and Eve thought that by obtaining one of the symbols of being a god, great knowledge, they too might be accepted into that group. Thus, two primary factors can be discerned which contributed to the violation of the norm: first, the desire of Adam and Eve to rise in the social structure and, second, the interaction with a member of a deviant culture who reduced their resistance to norm violation.

An alternative explanation of the reason why Adam and Eve desired to obtain knowledge has been suggested.[1] According to this view, they felt that by increasing their human capital (Becker 1967) they would raise their productivity and thereby positively affect their future income streams. However, since the processes of production and distribution were divorced in the Garden of Eden, the Protestant ethic not yet having emerged, this explanation seems to be spurious.

The last explanation of migration provided in the Bible is natural catastrophe. The case of the Great Flood is typical (Gen., chaps 6–9). Faced with death if a person does not move, he will accept migration, even when no clear-cut alternative location exists. However, in order to become part of the migration stream (or perhaps in this case migration ocean) the individual must accumulate sufficient resources to last the journey as well as have an acceptable means of transportation. In Noah's case, both of these first-order conditions were fulfilled, and therefore he and his household unit were able to migrate in response to the approaching catastrophe (Gen. 6.14, 15, 21). However, since the rest of humanity lacked the requisite first-order conditions, they were unable to migrate. Since this type of migration has no specific destination, it represents a classic and extreme form of migration in response to "push forces." On the other hand, the migration of the Jews from Egypt, as discussed earlier, could be viewed in terms of a broader application of the "push-pull" dichotomy, with the "pull" of higher welfare in the "Promised Land" being reinforced by the "push" of intolerable socioeconomic conditions in Egypt.

In addition to its implications concerning the nature of the migration decision process, this case study also provides considerable testimony concerning the processes of adaptation after migration. As was mentioned above, in Egypt the Jews did not assimilate due to the low economic and social class that they occupied in that country. Rather, the path of adaptation was one of economic pluralism and exploitation, with interaction between social groups restricted to the market place (Furnival 1948). After migration to the "Promised Land," the adaptation process was reversed for the Jews. In this case, rather than attempt to assimilate into the indigenous culture of Palestine, they proceeded to appropriate their land and cities (Deut., chaps. 1–4, and the Book of Joshua), in a way not unlike that used by the Europeans with respect to the American Indians in more recent times (Brown 1970). Thus, the indigenous population of Palestine was subjugated economically but, in this case, not socially. Little mention is made, but the implication is that after several generations the indigenous population was either absorbed through intermarriage or disappeared through attrition. In any event, the result was a coherent economic and social system defined by the culture of the recently migrated Jews.

---

[1] Many thanks to Dr Robert McIntyre for his help in this regard.

Although there are considerable problems in drawing modern parallels to the above-mentioned ancient processes, some parallels do indeed exist. Actually, a large portion of the modern literature on migration is merely a refinement of the processes described above. The nature of economic adaptation, however, has changed considerably since the time of Adam and Moses Due to the increased extent of economic and social stratification in most modern countries, the employment and role paths of the migrants have changed considerably from those of ancient times. Another important factor that has changed between the two time contexts is that the time perspective of the migrants and adapting postmigrants has become considerably shorter While in ancient times men were willing to accept current suffering and hardship so that future generations could benefit, modern man's time perspective is considerably shorter. Thus, adaptation has to have an immediacy that did not exist in the past. If substantial improvements in welfare do not result from migration within a short period after migration, then further migration or back migration may result This points to yet another difference between ancient and modern times that affects the migration-adaptation processes. Modern man is much less stable, geographically, in that migration over a considerable distance is often chosen more than once in one man's lifetime

In conclusion, while the "ancient" migration theories presented in this paper require some modification for them to be applicable to modern society, they contain considerable "Truth." Through the far more numerous opportunities for testing, these theories should be capable of both partial verification and useful extension and modification.

Robert Weinstein

Ohio State University

References

Becker, G. *Human Capital and the Personal Distribution of Income.* Ann Arbor· Univ. Michigan Press, 1967.

Brown, D. *Bury My Heart at Wounded Knee.* New York: Bantam, 1970.

Furnivall, J S *Colonial Policy and Practice.* Cambridge· Cambridge Univ. Press, 1948.

Book Reviews

Theories of Poverty and Unemployment. By David M. Gordon.
Lexington, Mass.: Lexington Books, 1972. Pp. ix + 179.

This book surveys and evaluates a large body of literature bearing on the personal distribution of earnings, with special emphasis on poverty Gordon argues that much of conventional theory on the subject is misplaced, that various distributional policies advocated under its perspective have not achieved their intended effects, and that empirical regularities uncovered by a different view of low-wage labor markets are inconsistent with received theory. The book consists of two parts. In the first, Gordon presents a brief, incomplete survey of what he calls orthodox theory, an exposition and review of the dual-labor-market hypothesis, and some explanation of recent radical theories of labor markets. The second part contrasts, compares, and criticizes the theories from both policy and empirical points of view, using existing empirical work as a guide

The book is organized around Thomas Kuhn's (1962) concept of a paradigm. Gordon evidently can recognize a paradigm when he sees one, for the book is the outcome of work initially supported by the Office of Economic Opportunity to stimulate "paradigm competition" regarding labor market behavior. This methodological paraphernalia, along with the tripartite distinction among radical, dual-market, and orthodox theories makes it tempting for a reviewer to draw some obvious analogies with political and religious ideology That temptation will be resisted, however, since the book is hardly polemical In any event, the competition is marred by the fact that the battlegrounds are not sharply drawn. An objective judge, should one exist in such matters, needs some criteria on which to choose a victor For that, the theories must be focused on a sufficiently interesting and similar set of problems for which their predictions differ. Though Gordon attempts the task, he really cannot accomplish it in any systematic manner, for the radical and dual-labor-market theories admittedly are far less complete than "orthodox" theory Indeed, they may be so incomplete as to make it problematical whether they deserve the label of paradigms, or even of theories for that matter Nevertheless, the book does succeed in stimulating interest in alternative approaches to labor economics. It contains an excellent review and bibliography of dual- and radical labor market literature and will serve as a relatively painless introduction for those curious to discover what all the shouting is about

What are the differences between the alternative theories? All are based on maximizing behavior, with varying emphasis on constraints Orthodoxy, as Gordon sees it, relates to the principles of marginal productivity and human capital. Dualists and radicals accept the lifetime-earnings emphasis of human capitalists as the empirical magnitudes of primary importance In addition, the former bring in some specific institutional features of labor markets, while the latter stress a concept of class interests as a major determinant of the institutional features. Both elevate the importance of specific job experience in generating lifetime productivity and earnings compared with the role of formal education as a productivity-augmenting device.

The fundamental insight of dual-labor-market literature is that the nature of work activities and job structure are endogenously determined, something that is masked over and often forgotten in writing enterprise production functions in terms of labor and capital aggregates. Profit-seeking employees structure work activities within their technological constraints to take advantage of existing supplies of personal characteristics of workers To this is added a kind of recursive feedback mechanism in which workers' tastes and job-related personal characteristics are affected, in an unspecified manner, by the nature of their work activities Finally, technological and product-demand conditions require a class of inherently unstable short-term jobs to serve buffer functions. The market becomes segmented between job types offering chances for advancement and large lifetime earnings and those offering high turnover and little in the way of future success. Tastes and worker characteristics evolve in ways that are consistent with this division, and the situation is maintained indefinitely Hence, the term "dual labor market" On this view, altering the lower tail of the earnings distribution involves literally breaking into the group holding unstable jobs

As a doctrinal matter, it is not clear why this theory should be labeled an alternative to orthodoxy, rather than an interesting and potentially important application of conventional concepts. Contributors to the dual-market literature apparently do not obtain their inspirations from conventional modes of thinking. Yet their differences in language may be misleading and perhaps not indicative *of real differences in substance What we seem to have here is a neat application of the theory of equalizing differences*, with the twist that both employers and workers can alter job and personal characteristics exchanged in the market I find it slightly scandalous that Gordon scarcely mentions the concept of equalizing differences in this connection Virtually none of the literature on the subject is reviewed or cited, of which Rottenberg's (1956) paper remains the best single exposition of its scope and limitations A strong case can be made for predicting that the major impact of dual-market and some radical literature on future labor market research will be to revive and extend the Smithian principles[1]

The radical position, revolving around the elusive concept of class, is far less easy to summarize. A central element is that employers not only structure jobs in their private self interest, but also with an eye toward thwarting class consciousness Worker choice is severely constrained, and focus naturally shifts from individual choices themselves to the constraining institutions The main difficulty with the class concept, aside from its unappetizing requirement of large-numbers collusion, is that no operational definition of class is available, a deficiency Gordon acknowledges at several points in the book Now collusion in the presence of government assistance, a favored theme of economists at the other end of the political spectrum, has more a priori appeal Radicals stress the interrelationships between public goods production and legislation that furthers dominant class interests and engenders taste formation consistent with these interests among other segments of the population. But all economic policies have income effects, some more blatant than others In fact, most of Gordon's more convincing illustrations of the uses of the class concept involve ex post evaluations of gains and losses to identifiable groups arising out of specific policies and legislation Ex post analysis is easy Out of all configurations of feasible institutions, can the radicals narrow the range of predictable winning coalitions that will obtain in practice?

Any theory of income distribution must face up to the fact that huge variation in earnings exists among individuals perceived to be identical within the limits of available data. The residual variance of earnings is at least as large as and probably larger than the systematic part. Independent of what interpretation is placed on the systematic components, controversies among competing theories of income distribution will remain unresolved until individual earnings can be predicted much more accurately than is possible at present. All the theories discussed here are highly deterministic in one sense or another. Perhaps the data are telling us something.

SHERWIN ROSEN

University of Rochester
and
Harvard University

References

Kuhn, Thomas S. *The Structure of Scientific Revolutions* Chicago: Univ Chicago Press, 1962

Rottenberg, Simon. "On Choice in Labor Markets" *Indus. Labor Relations Rev* (January 1956).

Soviet Wages Changes in Structure and Administration since 1956. By LEONARD J. KIRSCH

Cambridge, Mass M I.T. Press, 1972 $12 50

This volume is undoubtedly the most substantial published study of Soviet wages since Bergson's *The Structure of Soviet Wages* (1944) Although its focus is on the wage reform of 1958–60, it also traces the evolution of Soviet wage administration and wage structure (chiefly intraindustry differentials) through 1968

In the area of wage administration perhaps the most significant aspect of the reform has been the change in systems of wage payment Prior to the reform approximately two-thirds of Soviet industrial workers were paid according to straight or progressive piece rates By 1965 the latter had virtually disappeared while the former applied to less than one-quarter of industrial workers The single most important form of wage payment has become time rates plus premiums, while the piece rates that remain have increasingly become "collective" in nature (with the individual's earnings depending on the degree of fulfillment of his work team's collective output norm)

Kirsch's discussion of the rationale for the continuing Soviet reliance on some form of "payment by results"—whether premiums, piece rates, or some combination of the two—seems particularly useful It is clear that wage payment systems that stimulate a more intense work pace than would normally prevail under conditions of continuous full employment have much to recommend them But perhaps more important, "the wide application of various incentive systems adds flexibility to an otherwise relatively rigid system of wage administration" (p 43). In an administrative framework in which basic wage rates are determined by central authorities, the extensive use of incentive payment schemes administered by enterprise authorities provides a means of adapting earnings differentials to the requirements of local labor market conditions Enterprise authorities, in effect, can "correct" the centrally fixed basic rates to meet scarcities of particular skills via the implementation of incentive payment schemes

The most interesting portions of Kirsch's volume, at least for me, are those concerned with establishing recent trends in Soviet wage inequality The evidence amassed by Kirsch makes it clear that since the late 1950s, if not earlier, there has been "a significant reduction of wage inequality among industrial workers" (p 181). This trend reflects a reduction in the differentiation of earnings according to skill as well as a decline in interindustry wage differentials. Although Kirsch focuses on wage differentiation among industrial workers, a similar leveling tendency has also characterized the average earnings gap between workers, on the one hand, and "engineering-technical personnel," on the other. Thus, the decline in earnings inequality which Kirsch observes among workers may also apply to the broader category of "workers and employees"

These trends in the distribution of Soviet earnings are in striking contrast to the increased inequality which Henle has found for the distribution of earnings among wage and salary recipients in the United States in the period 1958–70 (*Monthly Labor Review* [December 1972], pp 16–27) It is curious that this trend should be treated by Henle as "perhaps almost inevitable in an advanced economy," while Kirsch interprets the opposite trend in Soviet Russia as a kind of natural accompaniment of increasing Soviet economic maturity (higher educational attainments, declining relative scarcity of skilled labor, a slowdown in the flow of peasant labor into industry). Is there a tendency to treat as "normal," as somehow inherent in a given stage of economic "maturity," whatever trends in wage structure we have discovered? Whatever explanations may be offered for the divergent current trends in Soviet and American wage structures, it seems probable that the kinds of labor market stratification which reinforce wage inequality in the United States are not an obstacle to the implementation of wage-leveling policies in the Soviet Union

Kirsch's findings for the period since 1956, when joined to the results of earlier studies in this area, now make it possible to trace the broad outlines of Soviet money wage differentiation from the early 1930s to the late 1960s But if, as Arthur Lewis once remarked, "socialism is about equality" (at least in part), it must be recognized that trends in Soviet money wage differentials are not always a reliable guide to broader changes in social and economic inequality under Soviet socialism The campaign against "equalitarianism" and the widening of money wage differentials during the early planning years were accompanied by the existence of rationing, widespread shortages of basic consumers' goods, and a substantial rise in communal consumption, all of which obviously limited the extent to which greater money wage differentials could be converted into increased real income differentials In more recent years the opportunities for those with relatively high money incomes to gain access to newly available goods and services (ranging from an increasing variety of consumer durables to occasional foreign travel) suggests that declining money wage differentials are compatible with at least some forms of increased social and economic differentiation.

These comments are not intended to detract from the value of Kirsch's study of the structure of Soviet money wages, but to indicate that the nature and forms of inequality in Soviet society remain fertile fields for continued study. The distribution of political power is the only dimension of Soviet inequality in which the important questions have long been settled.

Murray Yanowitch

Hofstra University

Impact of Uncertainty on Location. By Michael J. Webber.
Cambridge, Mass.: M.I.T. Press, 1972. Pp. xviii + 310.

It is well known that the introduction of uncertainty into economic models has enabled us to explain several phenomena imperfectly analyzed by models in a situation of certainty. In this book, Michael J. Webber sets out to discover how uncertainty affects location decisions.

In Webber's view, the main objective of location theory is to explain why economic activity is spatially concentrated. Starting from this position, the first part of the book (chaps. 2–4) is devoted to a critical analysis of existing theory in certainty situations. In the second part (chaps. 5–9), the author studies the role of uncertainty in the processes of the location of the firm.

The discrepancy observed in many cases between the ex post optimal location and the actual location has led some economists to reject the maximization of profit as an objective of the firm when making its location choice According to Webber, it is uncertainty which enables us to explain this discrepancy Indeed, the decision may be optimal given the firm's knowledge at the moment of decision but may no longer be so when the firm has obtained a better knowledge of its environment. Furthermore, the firm must sustain monetary costs in collecting information on possible location sites But the cost and the uncertainty may increase with distance This leads to a reduction in the number of locations considered by the firm. The same is true of the different markets where the firm may sell its output or buy its inputs

The author distinguishes between three types of uncertainty uncertainty about the behavior of rivals, about the state of the environment, and about production techniques A section of this book is devoted to each

The most original part of the work is undoubtedly the part devoted to uncertainty about rivals (chaps 6 and 7) Webber assumes that each firm making a location decision will not know its rivals' intentions, that is, the number, date, and place of future locations, and cannot change location Given the difficulty of analyzing all the possible processes, Webber proposes a very ingenious hypothesis the firm makes its decision while imagining that the other firms are going to locate at the same time He does not claim that all location decisions are taken simultaneously, but his hypothesis seems to enable us to simulate the decisions of firms which are expecting other firms to locate into the market at a later dâte. In this way the decision making may be analyzed within game models. It is assumed that the firms cannot communicate with each other and that they are seeking to maximize their profits safely in order to stay in business So the firms choose the least risky locations, they locate in existing towns mainly in the central town, whereas in a situation of certainty out-of-town locations would be possible. Thus one has a theory capable of explaining urban growth.

This stimulating conclusion is weakened once the specificity of the hypothesis is underlined It is supposed that the market is linear, and this leads to a highly simplistic representation of economic space In the process of location choice, production costs are unknown A demand spread over the market is partićularly applicable to consumer goods, but for enterprises producing intermediate or capital goods, the spatial structure of demand tends to be located at a finite number of points (Webber's analysis of wholesalers' location is not very satisfactory). The absence of information and coalition between firms is mainly true for small enterprises Indeed, if the head offices of large corporations are very often located in large towns in order to be near information sources and centers of political decision making, the location of production units is no longer made in the light of the same criterion.

Furthermore, the decision criterion used by Webber is not defined with enough precision. It corresponds in the majority of cases to a criterion of total uncertainty. about rivals. However, in the appendix to chapter 7 (p. 183) each firm is assumed to known its rivals' locations at the time it decides on its own optimal location. The game solution, if it exists, is given by the solution of the system of equations so obtained. The game bears a striking resemblance to Cournot's game, which is not a game in a situation of total uncertainty. Why this difference between the two?

Finally, Webber implicitly assumes the existence of an equilibrium in many cases (see, for example, p. 138), without analyzing the mathematical conditions necessary to ensure its existence. If we accept a more realistic representation of economic space, if the market area has any kind of shape, for example, or if demand is located at a finite number of points, or if the markets are spatially separated, the condition of convexity is not satisfied. In these cases, mixed strategies must be used, but Webber says nothing about them.

Chapters 8 and 9 discuss the impact of the other two forms of uncertainty about the state of environment and about production techniques. In Webber's view, they reinforce the tendencies toward spatial concentration. However, his argument is not rigorous enough. Furthermore, in chapter 8, Webber appears to be unaware of the contributions to the economic theory of the firm concerning the uncertainty of demand.

Finally, chapter 10 is devoted to a synthesis of his results and reflections about economic policy.

A better use of mathematical tools would undoubtedly have enabled Webber to reach a higher level of precision in his concepts and arguments If this had been the case, both the limits and the qualities of the model would have appeared more clearly The fact remains, however, that by his contribution Webber reveals several interesting areas for future research. I agree with him in his conclusion that "these results are sufficiently encouraging to indicate that an integrated mathematical statement of the general theory of the location under uncertainty should be made."

JACQUES THISSE

Fonds National de la Recherche Scientifique (Belgium)

International Capital Flows and Portfolio Equilibrium

Pentti J K. Kouri

International Monetary Fund

Michael G Porter

International Monetary Fund and Reserve Bank of Australia

This paper develops a model of international capital flows from a general equilibrium model of the financial markets of an open economy Capital flows are viewed as the mechanism by which a domestic excess demand for money is removed and consequently the key explanatory variables in the model are changes in domestic income, the current-account balance, changes in domestic monetary instruments, and changes in foreign interest rates since these will all affect either the demand for or supply of money The quarterly estimates of the capital flow equation for Australia, Italy, Germany, and the Netherlands appear to be consistent with the hypotheses derived from the model

This paper derives a model of international capital flows from a portfolio equilibrium model of an open economy under fixed exchange rates Our approach is a synthesis of the stock equilibrium approach to capital flows developed by Branson (1968)[1] and the monetary approach to balance-of-payments theory as developed for example in Mundell (1968) and

The views expressed in this paper are those of the authors and not necessarily those of the International Monetary Fund or the Reserve Bank of Australia Earlier drafts of this paper were presented at the winter meetings of the Econometric Society, Toronto, in December 1972, the meeting of the Project Link in Vienna in August 1972, and at a seminar at the Board of Governors, Federal Reserve System We have benefited from comments and suggestions from participants of these meetings, from our colleagues at the International Monetary Fund, and from a referee Mr Fernando Santos has provided valuable computational assistance

[1] This approach has been followed in numerous studies, for example, Lee (1969), Bryant and Hendershott (1970), Miller and Whitman (1970), Branson and Hill (1971), and Hodjera (1971)

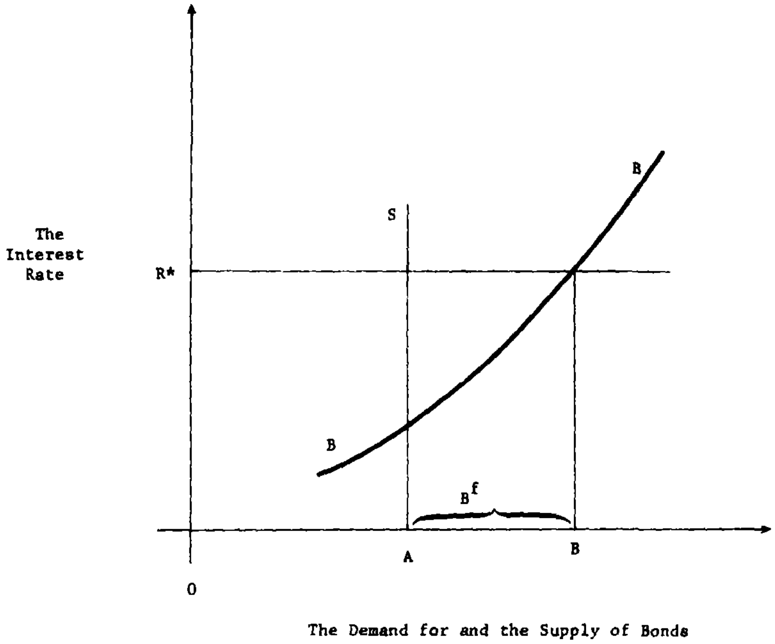


FIG 1 —The demand for foreign bonds when capital is perfectly mobile

interest rate, R^*, the domestic demand for bonds is OB The domestic net supply of bonds is OA The net stock of foreign bonds is equal to the difference between the total demand and the domestic supply (AB) In this framework, capital flows originate either from shifts in the BB schedule or from shifts in the net domestic supply (for instance, brought about by open-market operations) or from changes in the world interest rate R^* While this extreme case may not be empirically relevant yet (except when explaining capital flows within regions of a country), it would in our view be desirable to have a model that would contain this as a special case

A third problem with the Branson model is the fact that while it rests on solid microeconomic foundations it cannot easily be used to analyze many macroeconomic problems of importance, such as the effect of income fluctuations on the capital account, the link between the current account and capital flows, and the effect of changes in investment and saving behavior on the capital account In addition, the model treats the domestic interest rate as the instrument of monetary policy, thus implicitly assuming complete sterilization of the liquidity effects of capital flows We shall return to the question of sterilization later It suffices to point out here that it would be useful to be able to say something about the magnitude of, for instance, open-market operations needed to offset the liquidity effects of capital flows resulting from "independent" monetary policy.

A final comment relates to the monetary approach to the balance of payments The monetary theory of the balance of payments as developed, for instance, in Johnson (1971), explains the overall balance of payments as a function of excess demand for money. This approach is in contrast to much of the existing literature on capital flows which assumes implicitly that the demand for and supply of money determine the interest rate which then has an indirect effect on the capital account Similarly, while the monetary model implies a direct positive association between changes in income and the balance of payments, most of the capital flow models allow only for an indirect effect through the interest rate

In the next section we develop a simple model of capital movements that attempts to deal with the four problems discussed above

B A Portfolio Equilibrium Model

Our model is concerned with the mechanism through which equilibrium is obtained in the financial markets of a small open economy [4] It is assumed that changes in income, prices, and the stock of wealth are exogenously given, along with the associated current account of the balance of payments Thus monetary factors are assumed not to influence real variables in the current period Rather, it is assumed that changes in real variables and changes in the domestic component of the monetary base cause portfolio substitutions which lead both to capital flows and changes in the domestic interest rate By so constraining adjustment in the goods markets it is possible to derive capital-flow equations which facilitate relatively straightforward estimation and interpretation The model would appear reasonable when applied to quarterly data, given the abundance of evidence that monetary variables affect the real economy with a substantial and highly variable lag

For simplicity, it is assumed that there are only three kinds of financial assets in the economy base money, domestic bonds, and foreign bonds It is further assumed that expectations regarding both the price level and the exchange rate are stationary [5] The allocation of a given stock of financial wealth between these alternatives depends on the stock of wealth, income, the domestic interest rate, the foreign interest rate, and

---

[4] For a full exposition of a general equilibrium approach to monetary theory in a closed economy, see James Tobin (1969) A stock equilibrium approach to open monetary systems may be found in McKinnon and Oates (1966), Floyd (1969), and McKinnon (1969)

[5] In the empirical analysis we include dummy variables to capture the effect of speculation

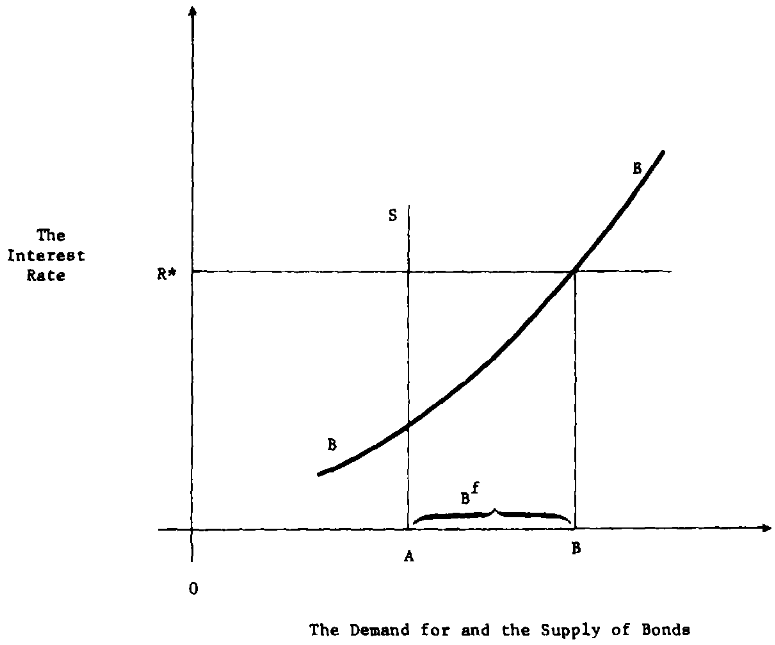


FIG 1 —The demand for foreign bonds when capital is perfectly mobile

interest rate, R^*, the domestic demand for bonds is OB The domestic net supply of bonds is OA The net stock of foreign bonds is equal to the difference between the total demand and the domestic supply (AB) In this framework, capital flows originate either from shifts in the BB schedule or from shifts in the net domestic supply (for instance, brought about by open-market operations) or from changes in the world interest rate R^* While this extreme case may not be empirically relevant yet (except when explaining capital flows within regions of a country), it would in our view be desirable to have a model that would contain this as a special case

A third problem with the Branson model is the fact that while it rests on solid microeconomic foundations it cannot easily be used to analyze many macroeconomic problems of importance, such as the effect of income fluctuations on the capital account, the link between the current account and capital flows, and the effect of changes in investment and saving behavior on the capital account In addition, the model treats the domestic interest rate as the instrument of monetary policy, thus implicitly assuming complete sterilization of the liquidity effects of capital flows We shall return to the question of sterilization later It suffices to point out here that it would be useful to be able to say something about the magnitude of, for instance, open-market operations needed to offset the liquidity effects of capital flows resulting from "independent" monetary policy.

A final comment relates to the monetary approach to the balance of payments The monetary theory of the balance of payments as developed, for instance, in Johnson (1971), explains the overall balance of payments as a function of excess demand for money This approach is in contrast to much of the existing literature on capital flows which assumes implicitly that the demand for and supply of money determine the interest rate which then has an indirect effect on the capital account Similarly, while the monetary model implies a direct positive association between changes in income and the balance of payments, most of the capital flow models allow only for an indirect effect through the interest rate

In the next section we develop a simple model of capital movements that attempts to deal with the four problems discussed above

B A Portfolio Equilibrium Model

Our model is concerned with the mechanism through which equilibrium is obtained in the financial markets of a small open economy [4] It is assumed that changes in income, prices, and the stock of wealth are exogenously given, along with the associated current account of the balance of payments Thus monetary factors are assumed not to influence real variables in the current period Rather, it is assumed that changes in real variables and changes in the domestic component of the monetary base cause portfolio substitutions which lead both to capital flows and changes in the domestic interest rate By so constraining adjustment in the goods markets it is possible to derive capital-flow equations which facilitate relatively straightforward estimation and interpretation The model would appear reasonable when applied to quarterly data, given the abundance of evidence that monetary variables affect the real economy with a substantial and highly variable lag

For simplicity, it is assumed that there are only three kinds of financial assets in the economy base money, domestic bonds, and foreign bonds It is further assumed that expectations regarding both the price level and the exchange rate are stationary [5] The allocation of a given stock of financial wealth between these alternatives depends on the stock of wealth, income, the domestic interest rate, the foreign interest rate, and other exogenous factors.[6] The basic model contains equilibrium conditions for the demand for and supply of money, domestic bonds, and foreign bonds The wealth constraint implies that only two equilibrium conditions are independent, since if excess demand for base money and domestic bonds is zero it follows that excess demand for foreign bonds is also zero These equilibrium conditions then determine the equilibrium stock of money, the domestic interest rate, and the stock of foreign bonds The basic model is specified table 1

---

[4] For a full exposition of a general equilibrium approach to monetary theory in a closed economy, see James Tobin (1969) A stock equilibrium approach to open monetary systems may be found in McKinnon and Oates (1966), Floyd (1969), and McKinnon (1969)

[5] In the empirical analysis we include dummy variables to capture the effect of speculation

The main feature of the model is that the income, wealth, and the expected-yield variables enter all demand functions for reasons set out in Tobin (1969). The partial derivatives of the respective arguments in the demand functions need little explanation However, it should be noted that a rise in income will, *ceteris paribus*, reduce the total demand for bonds since it increases the demand for money Equations (4), (5), and (6) are related via the wealth constraint (12) The money supply is the sum of the domestic and foreign assets of the central bank The monetary policy instrument, *NDA*, is assumed to change only as the result of open-market operations In empirical estimation, further refinements are necessary to allow for such factors as changes in reserve requirements.

The foreign component of the money supply changes either through capital flows or through the current account Capital flows are identically equal to the change in domestic holdings of foreign bonds minus the change in foreign holdings of domestic bonds Domestic bond market equilibrium is obtained when the net domestic stock of bonds issued by the government is equal to the total demand for domestic bonds, both from residents and foreigners [7] Finally, a simplification we have adopted is to express equilibrium conditions for base money rather than the total money supply, this being equivalent to assuming a money multiplier of one The essential features of the model are not substantially changed by incorporating a more complete banking system, however, we have refrained from such extensions in order to keep the model manageable

The domestic economy is assumed to be small, such that the foreign supply of foreign bonds is infinitely elastic at the foreign interest rate, R^*

---

[6] The stock of wealth is not determined within the model It changes from one period to another as a result of saving, which is exogenous to the model If, however, we allowed for changes in the valuation of existing assets, the value of the stock of wealth would become endogenous

[7] We abstract from the effect of budget deficits or surpluses on the supply of bonds We also ignore equity claims on real capital as well as holdings of physical commodities Furthermore, our concept of wealth is a narrow one in that it only refers to the market value of financial assets rather than the capitalized value of expected future income We also abstract from the problem of term structure of interest rates between different maturity classes of securities These abstractions should not affect the argument since our primary interest is in portfolio substitution between domestic and foreign securities and money.

TABLE 1

THE BASIC MODEL

	Equation No	Demand Functions	Partial Derivatives	Supply Relationships
Demand for base money	(4)	$Mo_D = L(Y, W, R, R^*, E)$	$L_Y, L_W > 0, L_R, L_{R^*} < 0$	
Net domestic demand for domestic bonds	(5)	$B_D = H(Y, W, R, R^*, E)$	$H_W, H_R > 0, H_{R^*} < 0, H_Y \lesseqgtr 0$	
Domestic demand for foreign bonds	(6)	$B_F = J(Y, W, R, R^*, E)$	$J_W, J_{R^*} > 0, J_R < 0, J_Y \lesseqgtr 0$	
Net foreign demand for domestic bonds	(7)	$B_D^* = F(Y^*, W^*, R, R^*, E)$	$F_R > 0, F_{R^*} < 0, F_{Y^*}, F_{W^*} \lesseqgtr 0$	
Total money supply	(8)			$Mo_s = NFA + NDA$
Domestic component of money supply	(9)			$\Delta NDA = -\Delta B_G$ (open-market operations)
Foreign component of money supply	(10)			$\Delta NFA = TC + CAB$
Total net capital inflow	(11)			$TC = \Delta B_D^* - \Delta B_F$
Domestic wealth constraints	(12)			$L(\ ) + H(\ ) + J(\ ) = W$
Money equilibrium	(13)			$Mo_D = Mo_S$
Domestic bond equilibrium	(14)			$B_D + B_D^* = B_G$

NOTE —Endogenous variables Mo_D = money demand, Mo_S = money supply, R = domestic interest rate, B_D = net domestic demand for domestic bonds, B_F = domestic demand for foreign bonds, B_D^* = foreign demand for domestic bonds, NFA = net foreign assets of central bank, TC = total net capital inflow Exogenous variables Y, Y^* = nominal income (domestic and foreign), W, W^* = nominal wealth (domestic and foreign), R^* = foreign interest rate, E = vector of risk factors, B_G = stock of government bonds held by private sector, NDA = net domestic assets of central bank, CAB = current account balance

However, the specific risk attached to domestic securities implies that the foreign demand curve for domestic bonds is upward sloping

Assuming that equilibrium is obtained within the period, the model can be solved for net capital flows and the change on the domestic interest rate as a function of the exogenous variables By invoking the wealth constraint (12), equation (5) and also equations (9) and (14) may be omitted, and hence the system can be reduced to the following two equations

$$\Delta L(Y, W, R, R^*, E) = \Delta NDA + CAB + TC, \qquad (15)$$

$$TC = \Delta F(Y^*, W^*, R, R^*, E) - \Delta J(Y, W, R, R^*, E) \qquad (16)$$

Invoking the wealth constraint, that is,

$$L_W + H_W + J_W = 1, \qquad (17)$$

and

$$L_i + H_i + J_i = 0, \qquad i = Y, R, R^*,$$

and solving, we obtain the reduced-form equations for TC and ΔR

$$\begin{aligned}\Delta R = \frac{-1}{H_R + F_R}\Big[&(H_{R^*} + F_{R^*})\,\Delta R^* + H_Y\,\Delta Y + H_W\,\Delta W \\ &+ CAB + \Delta NDA + F_{Y^*}\,\Delta Y^* + F_{W^*}\,\Delta W^*\Big]\end{aligned} \qquad (18)$$

$$\begin{aligned}TC = \frac{-1}{H_R + F_R}\Big\{&[(J_R - F_R)L_{R^*} + L_R(F_{R^*} - J_{R^*})]\,\Delta R^* \\ &+ [(J_R - F_R)L_Y - L_R J_Y]\,\Delta Y + (F_R - J_R)\,\Delta NDA \\ &+ (F_R - J_R)CAB + [(J_R - F_R)L_W - J_W L_R]\,\Delta W \\ &+ L_R F_{Y^*}\,\Delta Y^* + L_R F_{W^*}\,\Delta W^*\Big\}\end{aligned} \qquad (19)$$

In the limiting case of perfect capital market integration, we assume that H_R and J_{R^*} approach plus infinity while H_{R^*} and J_R approach minus infinity In other words, a small change in domestic or foreign interest rates generates infinite substitution of domestic and foreign bonds [8] In this case equations (18) and (19) are greatly simplified For example, the expression $(J_R - F_R)/[-(H_R + F_R)] \rightarrow 1$ as $J_R \rightarrow -\infty$, since $(J_R - F_R)/[-(H_R + F_R)] = (J_R - F_R)/(L_R + J_R - F_R) = [1 - (F_R/J_R)]/[(L_R/J_R) + 1 - (F_R/J_R)]$

---

[8] Strictly speaking we should note that there is a discontinuity in the partial derivatives As long as the initial portfolio contains both domestic and foreign bonds, the partial derivatives H_R, H_{R^*}, J_R, and J_{R^*} are infinite in the limiting case However, the total effect of an interest rate change on the holdings of, for instance, foreign bonds must be finite because of the wealth constraint even in the limiting case

The expression on the right-hand side of the above equation approaches one for finite values of L_R and F_R, as $J_R \rightarrow -\infty$ Taking the limiting values of the other coefficients gives us the following results under conditions of perfect capital market integration

$$\Delta R = \Delta R^* \frac{H_{R*} + F_{R*}}{-(H_R + F_R)}, \tag{20}$$

$$\begin{aligned} TC = {} & -\Delta NDA - CAB \\ & + \left(L_{R*} - \frac{J_{R*}L_R}{J_R}\right) \Delta R^* + L_Y \,\Delta Y + L_W \,\Delta W \end{aligned} \tag{21}$$

These results, in turn, may be further simplified if it is assumed that domestic and foreign securities offer equivalent risk, such that (J_{R*}/J_R) and (H_{R*}/H_R) approach minus one [9] In this case equation (20) becomes $\Delta R = \Delta R^*$ and the coefficient on ΔR^* in (18) is $L_{R*} + L_R$

C *Interpretation of the Capital-Flow Equation*

The reduced form capital-flow equation (19) and its limiting case (21) contain as arguments the monetary policy instrument, ΔNDA, the current account balance, CAB, the change in the foreign interest rate, ΔR^*, the change in domestic income and wealth, ΔY and ΔW, and the change in foreign income and wealth, ΔY^* and ΔW^* The extent to which monetary policy is offset by capital flows is estimated by the coefficients on ΔNDA and CAB, and the coefficients should be similar The limiting case of perfect mobility implies a coefficient close to minus one It is worth noting that in this limiting case the Branson model provides estimates of the interest elasticity of capital flow which are biased toward zero

Fluctuations in money demand caused by income are a source of capital flows, in the limiting case the income coefficient is L_Y, the inverse of the marginal income velocity of base money Foreign interest rate changes need not directly be a major source of fluctuations on the capital account as long as the domestic interest rate is allowed to adjust In the limiting case the coefficient on the foreign interest rate measures the interest elasticity of the demand for money Indirectly, changes in the foreign interest rate can have a much bigger effect on the capital account through the effect on income and the current account

---

[9] See "International Interest Rate Rations as Behavior towards Foreign Exchange Risk and Exchange Rate Expectations," appendix to Porter (1971) In the remainder of this paper we assume that as H_R and J_{R*} approach plus infinity and H_{R*} and J_R approach minus infinity, that H_R/H_{R*} and J_R/J_{R*} both approach minus one This amounts to assuming that securities offer equivalent risk in this limiting case

D. *Exchange-Rate Expectations*

It is clear that from time to time there are substantial capital flows arising from expectations of movements in the exchange rate—and this aspect of behavior has been neglected in our model A difficulty with developing an adequate model of exchange-rate expectations within our framework is that while an adaptive expectations model may explain expectations within the band, such a model will be of little relevance when the spot rate is at the ceiling or the floor, and there is an expectation of a shift in the parity Given this discontinuity in behavior, the most reasonable approach for our purposes would appear to be to use dummy variables for these periods when there were definite expectations of parity changes (e g , at the time of the 1969 revaluation of the Deutschmark)

II. Empirical Analysis

A The Estimating Equation

The ideal approach would be to simultaneously estimate the complete set of structural equations developed in the previous section. With the use of the a priori constraints all of the structural parameters could be identified This approach contains, however, several difficulties, a major problem being the unavailability of reliable data on the various stocks of assets and interest rates needed for simultaneous estimation. For instance, stock data on foreign assets and liabilities do not exist for most countries A second problem is the need for considerable institutional detail to adapt the theoretical framework to the peculiarities of any given country Third, there is the problem referred to previously that, as assets become perfect substitutes, individual structural equations cease to exist as well-defined functions Thus when domestic and foreign bonds become perfect substitutes there is no stable demand function for foreign bonds—the demand for foreign bonds becomes equal to the excess of aggregate demand for bonds over domestic supply of bonds As we show in the preceding sections, the reduced-form equations can be readily defined and interpreted even in this limiting case

The above considerations suggest the alternative approach of estimating the reduced-form equation of the previous section directly This approach provides the critical information on most of the questions that we are primarily interested in, such as the effect of income changes and monetary policy on the capital account and the relationship between capital flows and current account deficits or surpluses

Thus we estimate the following capital-flow equation.

$$TC = a_0 + a_1\,\Delta Y + a_2\,\Delta R^* + a_3\,\Delta NDA + a_4\,CAB$$
$$+ a_5\,\Delta E + a_6\,\Delta W + a_7\,\Delta Y^* + a_8\,\Delta W^* + u$$

The variables have the same meaning as in the previous sections Most of these variables are readily available in statistical form, the exceptions being the wealth and foreign income data, ΔW, ΔY^*, and ΔW^* There are no data available on W or W^* It is also difficult to construct a quarterly series on world income After some experimentation we decided to drop these three variables Since ΔW, and ΔY^*, and ΔW^* affect capital flows in opposite directions, and since we include the constant term, it seems reasonable to assume that the elimination of these variables will not greatly affect the results Furthermore, we know from the model that when the degree of capital mobility is high the coefficients a_7 and a_8 tend to be very small approaching zero in the limiting case Regarding the other coefficients we know that according to our model $a_1 > 0$, $a_2 < 0$, $-1 \leq a_3 \leq 0$, $-1 \leq a_4 \leq 0$ This provides a basis for a one-tail test of the significance of these coefficients Furthermore we would expect that $a_3 = a_4$, and we also know that both of them approach minus one as capital becomes perfectly mobile The ΔE variable measures changes in exchange-rate expectations As we discussed previously, we replace ΔE by dummy variables whenever appropriate

As has been noted above, all those components of the capital account that can be regarded as autonomous should be added to the current account Thus whenever the data are available we add official capital flows to the right-hand side In the case of Germany we also estimate the model on the assumption that long-term capital flows are exogenous

B Sterilization Policies and Possible Bias

The capital-flow equations estimated below treat the change in the net domestic assets of the central bank (ΔNDA) as exogenous However, it is possible that the authorities try to adjust ΔNDA at times so as to offset the liquidity effect of payments deficits or surpluses Indeed Michaely (1971) provides evidence that partial sterilization was the normal policy followed by a number of industrial countries in the period 1950–66 It is also possible that the monetary authorities assign monetary policy to the balance-of-payments target In this case the Central Bank would increase rather than decrease ΔNDA in response to an inflow of capital These practices give rise to a possible bias when we estimate the above equation by OLS It can be shown that in the case of sterilization the offset coefficient is biased toward minus one, while in the other case it is biased toward zero This can be shown in the context of the following simple model $TC = a = b\,\Delta NDA + cX_1 + u$, $\Delta NDA = \alpha + \beta TC + \gamma X_2 + v$, where X_1 is an exogenous influence on TC, X_2 some other target of monetary policy, and u and v are the error terms A negative coefficient on TC in the second equation implies sterilization For example, if $\beta = -1$, then there is complete sterilization With the usual assumptions regarding the error terms we can derive the asymptotic bias of the OLS estimate of the offset coefficient ($\hat{b}$) It can be shown that the bias on $\hat{b}$ is given by the formula[10] plim $(b - \hat{b}) = [\beta(1 - b\beta)\sigma_u^2]/(\gamma^2\bar{M}_{x_2x_2} + \sigma_v^2 + \beta^2\sigma_u^2)$, where σ_u^2 = the variance of u, σ_v^2 = the variance of v, and $\bar{M}_{x_2x_2}$ = the asymptotic variance of X_2 This formula shows first of all that the bias is negative when β is negative (as long as $b\beta$ is less than one, which it must be for stability) Thus $\hat{b}$ is biased toward minus one. The critical question is the magnitude of this bias, and that depends on the ratio of the variance of the error term in the capital-flow equation, (σ_u^2), to the variance of X_2, and the variance of the error term in the sterilization equation, (σ_v^2) This means that so long as the capital-flow equation is well specified and so long as discretionary changes in monetary policy as opposed to automatic sterilization dominate ΔNDA the bias is likely to be small It is important therefore to include dummies to capture large speculative capital movements, such as occurred in Germany Otherwise the sterilization which took place and was reflected in ΔNDA would seriously bias the offset coefficient toward minus one[11]

A related question to that discussed above is the fact that the Central Bank may not directly control certain components of *NDA*, such as the level of commercial bank borrowing from the Central Bank There is little that can be done about this short of a comprehensive model of the monetary sector It is not unreasonable to assume that the Central Bank can control *NDA* if it so decides, but whether it has always done so is quite a different question We shall maintain the hypothesis that ΔNDA can be taken as an approximate measure of monetary policy

III. The Results for Germany, Australia, Italy, and the Netherlands

A A Brief Summary

The estimates of the capital-flow equation for Germany, Australia, Italy, and the Netherlands are reported in tables 2 and 3 and the behavior of the actual and fitted values is illustrated in figures 2–5 Description of the variables and the data sources are provided in Appendix II The starting point of the estimation period has been largely determined by data

---

[10] For an analysis of the simultaneous-equations bias, see for instance, Johnston (1972), chap 12

[11] A recent paper of Argy and Kouri (1972) attempts to isolate the offsetting and sterilization effects by a two-stage procedure Their results for Germany, Italy, and the Netherlands indicate partial sterilization and offsetting of similar magnitude to what we obtain here The problem with econometric analysis of Central Bank behavior is that in most cases discretionary changes are likely to dominate automatic and regular response patterns As argued above, this tends to reduce the bias when the policy measures are treated as exogenous

TABLE 2

Estimates of the Capital-Flow Equation for Germany

Dependent Variable	Constant	ΔY	ΔR^*	ΔNDA	ΔRR	$\Delta NDA - \Delta RR$	CAB	$CAB + LTC$	$\Delta D1$	$\Delta D2$	$\Delta D3$	$\Delta S4$	$\bar{R}^2$	SE	D-W
(1) *TC*	243 60	0 11	31 06	−0 74	0 86		−0 93		1,772 00	1,909 00	3,074 00	1,751 00	960	818 00	2 097
	(1 32)	(4 02)	(0 12)	(12 47)	(8 14)		(9 99)		(2 86)	(2 87)	(3 04)	(6 94)			
(2) *TC*	238 50	0 11	115 40			−0 77	−0 95		1,798 00	2,027 00	2,703 00	1,783 00	960	816 00	2 172
	(1 30)	(4 11)	(0 45)			(18 40)	(10 58)		(2 91)	(3 11)	(2 92)	(7 16)			
(3) *STC*	252 10	0 12	75 78	−0 72	0 86			−0 94	1,797 00	1,892 00	3,061 00	1,693 00	937	822 00	2 006
	(1 36)	(3 84)	(0 29)	(9 55)	(8 08)			(8 57)	(2 80)	(2 61)	(2 90)	(5 56)			
(4) *STC*	250 90	0 11	144 50			−0 77		−0 99	1,876 00	2,087 00	2,632 00	1,780 00	937	820 00	2 113
	(1 36)	(3 79)	(0 57)			(15 25)		(9 91)	(2 95)	(3 03)	(2 81)	(6 19)			

Note —The sample period = 1960 1–1970 4, figures in parentheses = *t*-statistics, $\bar{R}^2$ and SE are adjusted for the degrees of freedom, D-W = Durbin-Watson statistic The variables are defined as follows *TC* = total private capital flows *STC* = total private short term capital flows *LTC* = total private long-term capital flows, *CAB* = current account balance including official capital flows, ΔY = change in income at an annual rate, ΔR^* = change in the 3-month Euro-dollar rate, $\Delta ND4$ = change in the net domestic assets of the Central Bank, ΔRR = policy determined change in required reserves *D*1, *D*2, and *D*3 are speculative dummies They are equal to one in 1961 1 1968 4, and 1969 3, respectively, and zero elsewhere *S*4 = 1 in the fourth quarter and zero otherwise *TC*, *STC*, *LTC*, *CAB*, *Y*, and *ND4* are measured in millions of Deutschmarks All data are seasonally unadjusted

TABLE 3

ESTIMATES OF THE CAPITAL-FLOW EQUATION FOR AUSTRALIA, ITALY, AND THE NETHERLANDS

Country	Dependent Variable	Constant	ΔY	ΔR^*	ΔNDA	CAB	$DAUST$	$\bar{R}^2$	SE	D-W
(1) Australia	*TC*	78 16	0 057	−11 84	−0 47*	−0 63	216 20	816	63 90	1 875
		(4 11)	(2 84)	(0 79)	(5 29)	(6 74)	(6 08)			
(2) Italy .	*TC*	84 84	0 027	−51 35	−0 43	−0 98		665	132 00	2 554
		(1 25)	(6 26)	(1 22)	(4 36)	(4 95)				
(3) The Netherlands	*TC*†	88 98	0 021	−94.09	−0 59	−0 98		820	193 00	2 539
		(2 91)	(5 01)	(1 62)	(7 58)	(10 88)				

NOTE —The sample periods are 1961 3–1972 3 for Australia, 1964 1–1970 4 for Italy and 1960 1–1970 4 for the Netherlands The variables are measured in millions of Australian dollars, billions of Italian lire, and millions of Dutch guilders, respectively *DAUST* is a speculation dummy for Australia *It is equal to zero up to 1970 4 and one thereafter* For other definitions, see table 2

* ΔNDA is adjusted for changes in the level of restricted deposits

† Official capital flows are included in *TC* and not in *CAB*

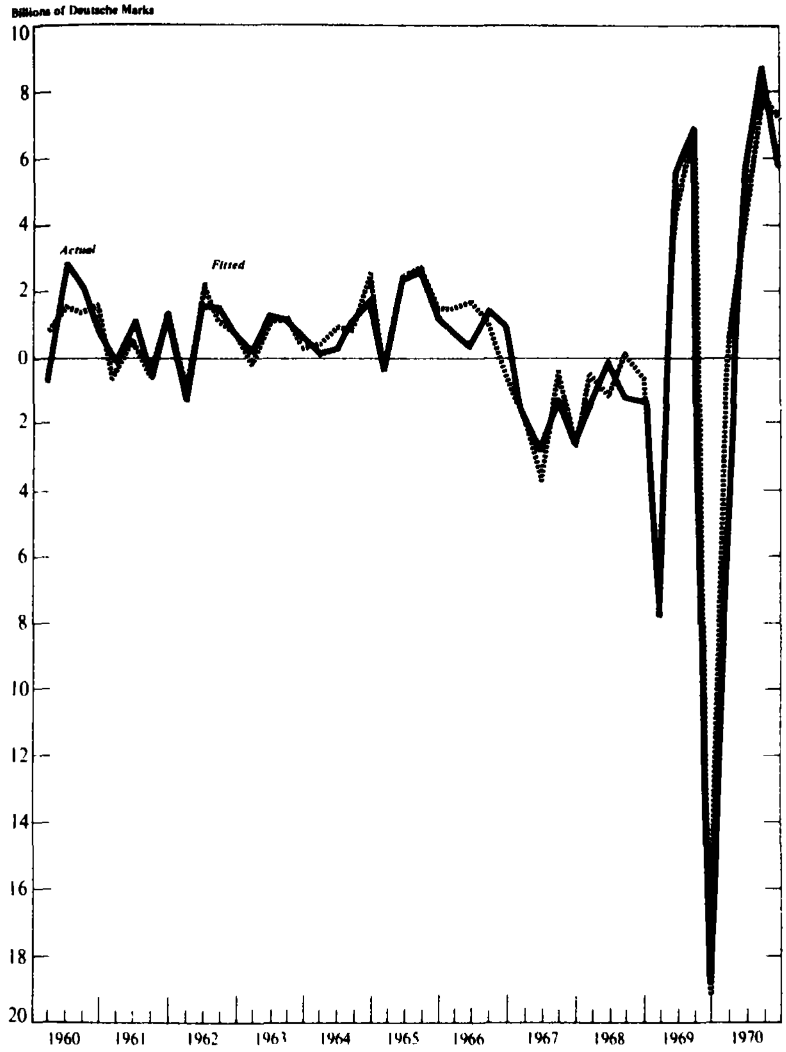


Fig 2 —Germany total private capital flows, quarterly 1960 1–1970 4 (source table 2, eq [2])

availability The terminal point was dictated by the flotation of exchange rates in early 1971, except in the case of Australia

A striking feature of the results is that the change in income is found to be significant in all four cases, suggesting that quarterly changes in income create changes in money demand which are at least partially satisfied by capital flows The offset coefficient, which measures the

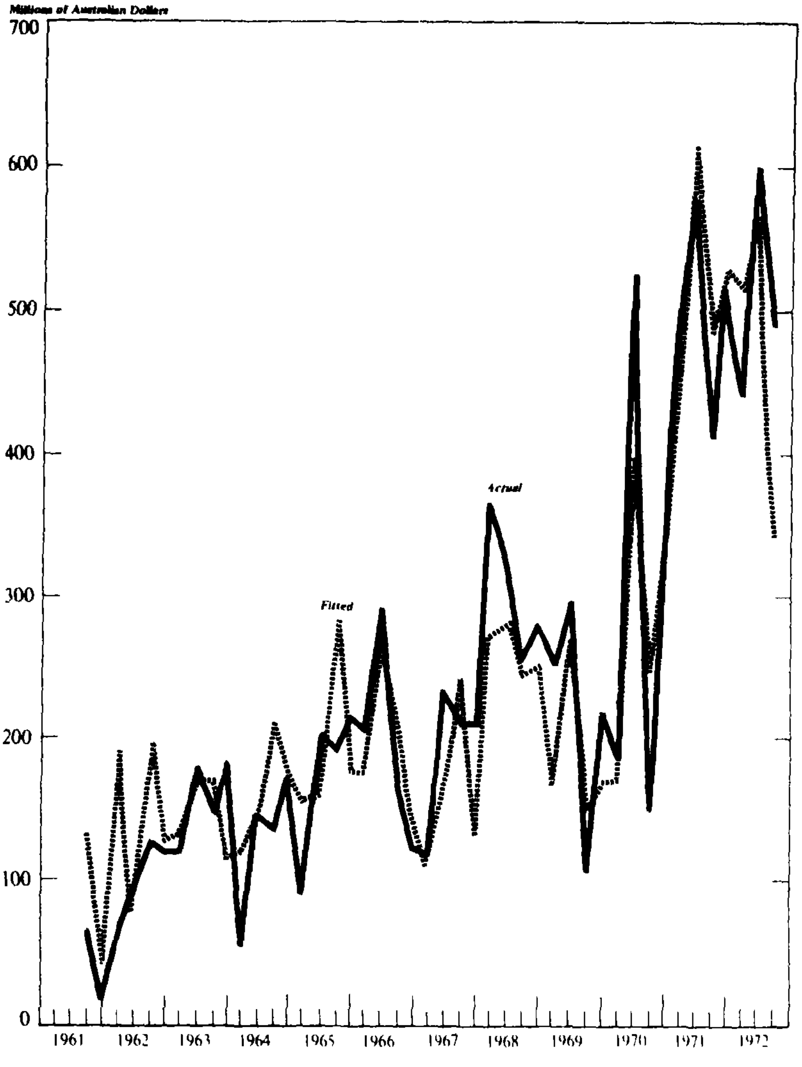


FIG 3 —Australia total private capital flows, quarterly 1961 3–1972 3 (source table 3, eq *[1]*)

extent to which capital flows offset policy-induced changes in the monetary base, is statistically significant in all cases The point estimates are −0 77 for Germany, −0 59 for the Netherlands, −0 47 for Australia, and −0 43 for Italy All of these estimates are statistically different from minus one, which suggests that sterilization is possible in these countries, at least in the short run, if the Central Bank is willing to accumulate (or lose) reserves Similarly the offset coefficients on the current account are

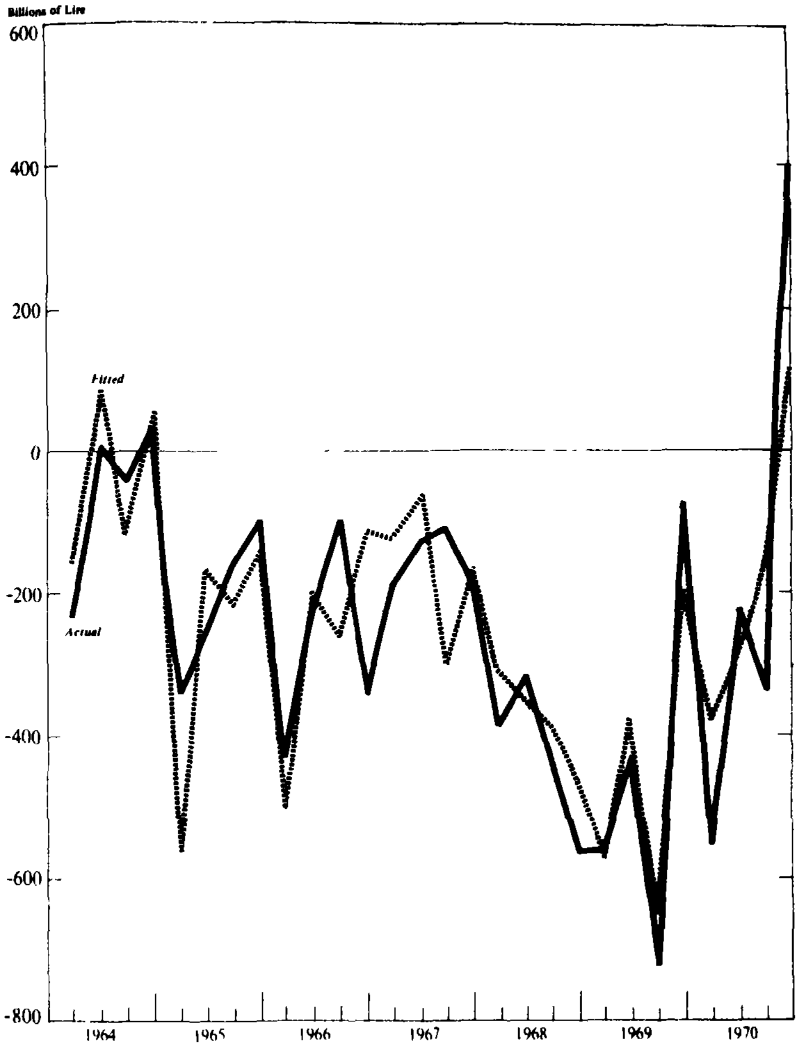


FIG 4 —Italy total private capital flows, quarterly 1964 1-1970 4 (source table 3, eq [2])

all statistically significant and very high, −0 99 for Germany, −0 98 for Italy, −0 98 for the Netherlands, and −0 63 for Australia In the German and Australian case the two offset coefficients are not significantly different, while for Italy and the Netherlands the offset coefficient on the current account is substantially higher We have no good explanation available for these differences

A disappointing feature of the results is the fact that the change in the

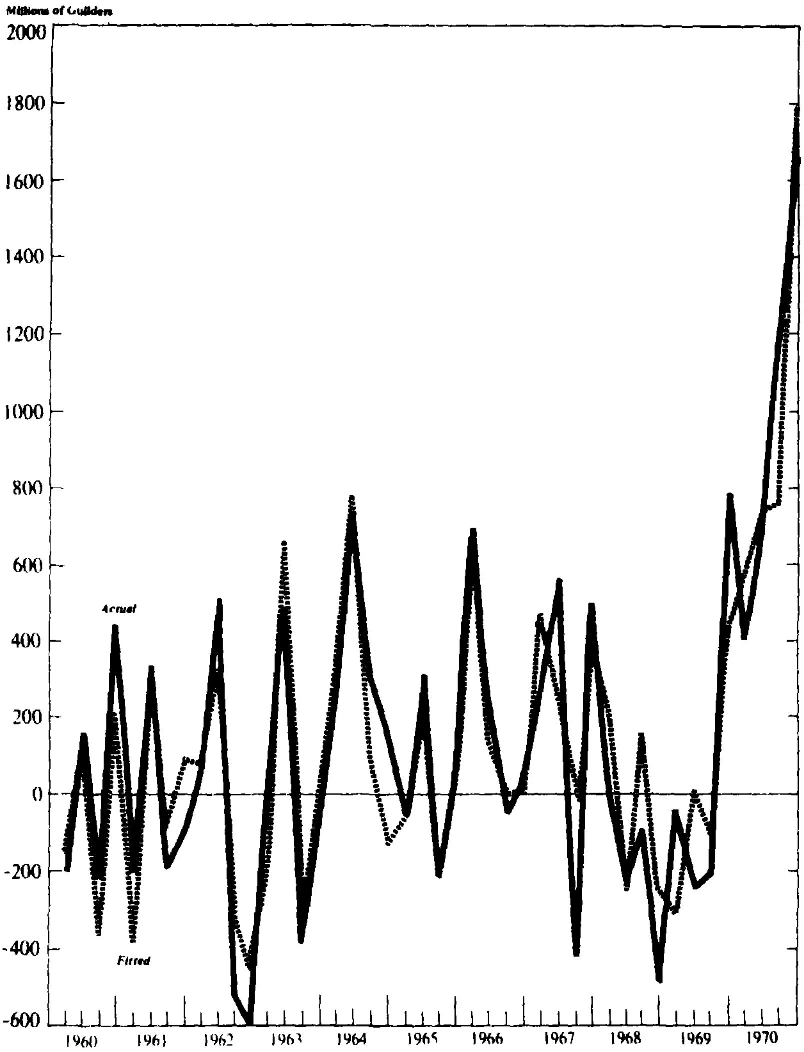


FIG 5 —Netherlands total capital flows, quarterly 1960 1–1970 4 (source table 3, eq [3])

Euro-dollar rate is statistically insignificant in all cases Although it should be remembered that the coefficient on the foreign interest rate does not measure the interest sensitivity of capital flows (which is measured by the offset coefficient), one would still expect a nonzero coefficient on the Euro-dollar rate as long as the demand for money is interest sensitive. There are two possible explanations for this result One is that the Euro-dollar rate may not in all cases be the relevant foreign interest rate The other and probably the more important reason is the fact that we have not satisfactorily incorporated the forward market and exchange-rate expectations into our model. Since the forward premium, in particular in the case of Germany, has fluctuated considerably and often in a direction opposite to the Euro-dollar rate, it is possible that its exclusion tends to suppress the interest rate coefficient. Speculation dummies turned out to be significant in the case of Germany and Australia while no dummies introduced on a priori grounds were statistically significant in the case of Italy and the Netherlands.

The overall explanatory power of the equations is quite satisfactory The Durbin-Watson statistic indicates no serial correlation in the case of Germany and Australia while it falls in the inconclusive region in the two other cases. A short discussion of the individual country results is provided in the next four sections.

B. Germany

Two separate policy instruments were incorporated into the equation the policy-induced change in average required reserves (ΔRR) and the change in the Bundesbank's net stock of domestic assets (ΔNDA) The difference of these two corresponds to the domestic component of the extended monetary base The other variables are the change in income, the change in the 3-month Euro-dollar rate, the current account balance plus official capital flows, and a number of dummies designed to capture the effect of speculative capital flows In addition, we include a dummy which takes care of the pronounced end-of-year capital inflow, and a subsequent outflow

As is shown in table 2, the offset coefficients on the two monetary instruments are both very high, 0 74 for ΔNDA and 0 86 for ΔRR[12] As expected these coefficients are not significantly different from each other When we combine them we obtain an offset coefficient of 0 77 on the composite monetary instrument ($\Delta NDA - \Delta RR$) with a standard error of 0 04 While the coefficient is high, it is significantly different from one (for brevity we do not always mention the sign, it should be remembered that all the offset coefficients are negative) This means that the Bundesbank can exercise some degree of monetary independence in the short run if it is willing to accept the reserve movements which that policy entails This point is illustrated by the following example Suppose that the Bundesbank wants to decrease the monetary base by 100 million marks This would require a reduction of 435 million marks in NDA, and it would imply an increase in foreign exchange reserves of 335 million marks

---

[12] These results are similar to those obtained earlier by Porter (1972) by using monthly data

The offset coefficient on the current account is also significant and close to one It is not, however, statistically different from the offset coefficient on (ΔNDA – ΔRR) The high value of the coefficient on *CAB* suggests that private capital flows have to a large extent financed fluctuations on the current account Thus the behavior of the overall balance of payments is dominated by the factors that influence capital flows. Monetary policy is one of these factors Income fluctuations are another The income coefficient indicates that an increase in income at an annual rate of 1 billion marks will generate capital inflows of some 110 million marks The income variable explains partly the clear procyclical pattern of the overall balance of payments The other reason for this is the countercyclical behavior of monetary policy as reflected in the movements of ΔNDA

The Euro-dollar rate is not significant in any of the equations We tend to think that this is the result of the inadequate treatment of exchange-rate expectations In future research it would be important to incorporate the forward market into the model to tackle exchange-rate uncertainty more directly than has been possible in this study

Three dummies introduced on basis of a priori knowledge of speculative episodes turned out to be significant The first dummy captures the speculative inflow and outflow surrounding the March 1961 revaluation of the Deutschmark The second dummy takes care of the speculative inflow in the fourth quarter of 1968, which was due to an expectation of a revaluation of the Deutschmark The third dummy captures the inflow and outflow surrounding the second revaluation of the Deutschmark in October 1969

The magnitude of the last dummy appears rather small compared with the magnitude of capital movements in the last two quarters of 1969 One reason for this may be that the speculative inflow extended over a longer time period than the outflow just after the revaluation The other reason is the fact that if the Central Bank does not attempt to sterilize the liquidity effect of inflow (by changing ΔNDA) the inflow will suppress domestic interest rates, which will have an opposite effect on the net magnitude of capital flows The coefficient on the speculation dummy measures the combined effect of these two factors

Finally, equations (6) and (7) report the results that we obtain by treating long-term capital flows as an exogenous variable The results are remarkably similar, indicating that in the shortrun the adjustment takes place largely through short-term flows An examination of the data also reveals that the variation on the long-term capital account is substantially less than that on the short-term account.

C. *Australia*

Australia has experienced sustained private capital inflow over the period studied, 1961–72, and there have been substantial variations in the composition and magnitude of these flows. Our model does reasonably well in explaining these movements. The offset coefficient on ΔNDA, which has been adjusted for changes in required reserves (statutory reserve deposits), is 0.47, while that on the current account is 0.68. They are not statistically different, and both are substantially less than the coefficients we obtained for Germany. The income variable is also significant, a rise in income of $1 billion Australian ($A) generating capital inflows of $57 million. The Euro-dollar rate is again insignificant although it has the correct sign. We include a dummy variable to allow for speculative inflows in 1971 and 1972. The coefficient of $A216 million suggests that speculative inflows in this period amounted to a total of some $A1.5 billion (see table 3).

D. *Italy*

Capital flows, typically outflows, from Italy over this period have varied considerably. The model appears to explain these variations quite well, capturing all the major swings during the period studied. All variables except the Euro-dollar rate are significant. The offset coefficient on ΔNDA is 0.43 with a standard error of 0.099. The offset coefficient on the current account, also significant, is much higher, being close to one. This is in contradiction with what we would expect from the model. The income variable is also significant. It is interesting to note that Branson and Hill (1971) found that the gap between potential and actual GNP is highly significant in explaining Italian capital flows in the framework of their model. In addition, we experimented with a number of speculative dummies, but these turned out to be insignificant.

E. *The Netherlands*

The results for the Netherlands are similar to those obtained for Australia and Italy. The offset coefficient is 0.58 with a standard error of 0.078. The offset coefficient on the current account is again close to one and statistically greater than that on ΔNDA. In terms of the model the sharp increase in net capital inflow in the last quarter of 1969 and in 1970 can be explained by the rapid increase in income, large deficits on the current account, the decline in the domestic component of the monetary base, and a fall in the Euro-dollar rate in 1970.

IV. Conclusions

A most interesting aspect of our results is that changes in income are highly significant in explaining capital flows in all cases This provides confirmation of the view that the behavior of capital flows reflects divergent growth rates and cyclical phases between countries—with capital flows accommodating the resultant fluctuations in money demand Accordingly, while it may appear that capital movements largely respond to changes in interest differentials between countries, a closer analysis reveals that the more basic sources of capital flows are fluctuations in money demand and supply, with income fluctuations being important in explaining shifts in money demand.

Our results demonstrate that for all four countries capital flows are to a large part the result of changes in monetary policy, as reflected in changes in the net domestic assets of the Central Bank and the required reserves The current-account balance enters our capital-flow equations as an exogenous source of change of base money We find that the current-account balance tends to induce offsetting capital flows, thereby stabilizing the balance of payments

This means that the behavior of the overall balance of payments is dominated by factors that influence capital flows, that is, changes in income (with a positive sign), changes in the world interest rate, changes in exchange-rate expectations, and changes in monetary policy instruments

Our results also indicate that while monetary policy has a strong effect on the capital account in all the four countries there is still some room for independent monetary policy, if only in the shortrun, if the Central Bank is willing to tolerate the large reserve movements which that policy entails.

Appendix I

The Bias in the Branson Model

Consider a simplified version of the Branson model: $\Delta B^F = a_0 + a_1 \Delta R + a_2 Z_1 + u$, where ΔR is the change in the domestic interest rate, Z_1 is an exogenous variable explaining capital flows, and u is the error term Suppose that capital flows affect interest rates, that is, $\Delta R = b_0 + b_1 \Delta B^F + b_2 Z_2 + v$, where Z_2 is an exogenous variable explaining the changes in ΔR, and v is the error term. We make the usual assumptions regarding the error terms. It can be shown that the asymptotic bias of the OLS estimate of a_1 is given by [13] plim $(\hat{a}_1 - a_1) = [b_1(1 - a_1 b_1)\sigma_u^2]/(b_2^2 \overline{M}_{z_2 z_2} + \sigma_v^2 + b_1^2 \sigma_u^2) < 0$, where $\overline{M}_{z_2 z_2}$ = asymptotic variance of Z_2, assumed to be finite, σ_u^2 = variance of u, and σ_v^2 = variance of v. It is assumed that the asymptotic covariances between Z_1, Z_2, u, and v are zero.

---

[13] See Johnston 1972, chap 12. A detailed analysis of sources of least-squares bias is given in Bronfenbrenner (1953)

To obtain some definite conclusions regarding the magnitude of the bias it is useful to note that as the degree of capital mobility, as measured by a_1, approaches infinity, the absolute bias approaches infinity What this means is that the higher the degree of capital market integration the less appropriate is the OLS estimate of interest elasticity of capital flows, since the estimated coefficient becomes severely downward biased Furthermore, the OLS estimate of a_2 is biased downward. Its asymptotic bias is given by: plim $(\hat{a}_2 - a_2) = (b_1^2\sigma_u^2)/(b_2^2\bar{M}_{z_2z_2} + \sigma_v^2 + b_1^2\sigma_u^2) < 0$. This bias does not depend on a_1, hence for large values of a_1 the bias on $\hat{a}_1$ is larger than that on $\hat{a}_2$

Appendix II

Symbols and Sources of Data Used in Regressions

TC = total private capital flows (including official capital flows for the Netherlands), from IMF data fund,

STC = total short-term private capital flows (including official capital flows for the Netherlands), from IMF data fund,

CAB = current account balance (including official capital flows for Germany, Australia, and Italy), from IMF data fund,

R^* = 3-month Euro-dollar rate, from the *International Financial Statistics* (*IFS*),

Y, Germany = mass income at an annual rate, from the Bundesbank's monthly reports (BMR),

Y, Australia = national income, from *IFS*,

Y, Italy = GDP at market prices, generated from leading indicators reported in the *OECD Main Economic Indicators*,[14]

Y, the Netherlands = GDP at market prices, generated from leading indicators reported in the *OECD Main Economic Indicators*,

NDA, Germany, Italy, and the Netherlands = net domestic assets of the Central Bank, from the *IFS*,

NDA, Australia = net domestic assets of the Central Bank less restricted deposits, from the *IFS*,

ΔRR, Germany = change in required reserves. Computed according to the formula $\Delta RR = (r - r_{-1})L_{-1}$, where r is the average required reserve ratio and L is the total stock of bank liabilities subject to reserve requirements, from BMR

[14] Since there are no quarterly data on income for Italy and the Netherlands, a quarterly GDP series was generated from some leading indicators available on quarterly basis For Italy the value of retail sales and the value of industrial production (industrial production index times the consumer price index) were used as indicators The regression of annual GDP on the annual averages of these two variables (*X* and *Y*, respectively) for the period 1960–70 yielded

$$\begin{aligned} \text{GDP} &= -5208.0 + \underset{(9.58)}{348.1X} + \underset{(1.56)}{0.712Y} \\ & \;\;(0.53) \\ \bar{R}^2 &= 999 \qquad \text{SE} = 307.0 \qquad \text{D-W} = 2.063 \end{aligned}$$

For the Netherlands the quarterly series on the value of investment was in addition used as an indicator The annual relationship between GDP at market prices and these indicators for the period 1960–70 is

$$\begin{aligned} \text{GDP} &= \underset{(0.94)}{2048.0} + \underset{(2.62)}{994.2X} + \underset{(2.33)}{166.9Y} + \underset{(3.87)}{237.3Z} \\ \bar{R}^2 &= 998 \qquad \text{SE} = 743.0 \qquad \text{D-W} = 2.606 \end{aligned}$$

The data on *X*, *Y*, and *Z* were obtained from the *OECD Main Economic Indicators*

References

Argy, V., and Kouri, P. J. K "Sterilization Policies and Volatility in International Reserves." Paper presented at the Wingspread Conference on "National Monetary Policies and the International Monetary System," Racine, Wisc , July 1972

Branson, W H *Financial Capital Flows in the U S Balance of Payments.* Amsterdam North-Holland, 1968

Branson, W H , and Hill, R D "Capital Movements in the OECD Area " *OECD Economic Outlook, Occasional Studies* (December 1971)

Bronfenbrenner, J "Sources and Size of Least-Squares Bias in a Two-Equation Model " In *Studies in Econometric Method,* edited by W C Hood and T C Koopmans Cowles Foundation Monographs, no. 14 New York Wiley, 1953

Bryant, R C , and Hendershott, P H *Financial Capital Flow in the Balance of Payments of the United States An Explanatory Empirical Study* Princeton Studies in International Finance, no 25 Princeton, N J Princeton Univ Internat Finance Sec., 1970

Floyd, J E "International Capital Movements and Monetary Equilibrium " *A E R* 59, no 4, pt 1 (September 1969) 472–92

Hodjera, Z "Short-Term Capital Movements on the United Kingdom, 1963–67 " *J P E* 79 (July/August 1971) 739–75

——— "International Short-Term Capital Movements A Survey of Theory and Empirical Analysis " *IMF Staff Papers* (forthcoming)

International Financial Statistics Various issues Washington Internat Monetary Fund

Johnson, H G "The Monetary Approach to Balance of Payments Theory " Unpublished monograph, 1971

——— *Inflation and the Monetarist Controversy* Amsterdam North-Holland, 1972

Johnston, J *Econometric Methods* New York McGraw-Hill, 1972

Leamer, E R , and Stern, R M "Problems in the Theory and Empirical Estimation of International Capital Movements " In *International Mobility and Movement of Capital,* edited by Fritz Machlup, Walter S Salant, and Lorie Tarshis New York Nat Bur Econ Res , 1972

Lee, C H "A Stock Adjustment Analysis of Capital Movements The United States–Canada Case " *J P E* 77 (July/August 1969) 512–23

McKinnon, R I "Portfolio Balance and International Payments Adjustment " In *Monetary Problems of the International Economy,* edited by R. A Mundell and A Swoboda Chicago· Univ Chicago Press, 1969

McKinnon, R I , and Oates, W E *The Implications of International Economic Integration for Monetary, Fiscal, and Exchange Rate Policy* Princeton Studies in International Finance, no 16. Princeton, N J Princeton Univ Internat Finance Sec , 1966

Michaely, M *The Responsiveness of Demand Policies to Balance of Payments Postwar Patterns* New York Nat. Bur Econ Res , 1971.

Miller, N C , and Whitman, M "A Mean-Variance Analysis of United States Long-Term, Portfolio Foreign Investment." *Q J.E* 84 (May 1970) 173–96

Mundell, R A *International Economics* New York Macmillan, 1968.

OECD *Main Economic Indicators.* Various issues Washington: Org. Econ. Cooperation and Development.

Polak, J. J. "Monetary Analysis of Income Formation and Payments Problems " *IMF Staff Papers* 6 (March 1958) 1–50

Polak, J J , and Argy, V. "Credit Policy and the Balance of Payments " *IMF Staff Papers* 18 (March 1971) 1–24.

Porter, M G. "A Theoretical and Empirical Framework for Analyzing the Term Structure of Exchange Rate Expectations" *IMF Staff Papers* 18 (November 1971): 613–45.

——— "Capital Flows as an Offset to Monetary Policy The German Case" *IMF Staff Papers* 19 (July 1972) 395–424

Tobin, J "A General Equilibrium Approach to Monetary Theory" *J Money, Credit and Banking* 1 (February 1969) 15–29

Taxation and the Rates of Return from Some U.S. Asset Holdings Abroad, 1960–1969

Herbert G. Grubel

Simon Fraser University

The paper shows theoretically that U S tax laws are designed to induce equalization of private rates of return from investment in the United States and abroad but that, unless the marginal government expenditures on investment equal tax payments by the investment, there is a difference between private and social rates of return to U S. foreign investment Calculations covering the period 1960–69 for developed countries and U S holdings of bonds, equities, and direct investment in manufacturing showed that private rates of return generally were positive but that social rates of return from investing abroad rather than domestically were strongly negative, assuming zero marginal government expenditures and disregarding all externalities

In the past the basic philosophy underlying U S tax laws and other regulations guiding international long-term capital flows have been to encourage the maximization of world income The policies flowing from this philosophy by and large enabled capital to move anywhere in the world where private owners could earn highest returns[1]

An important question facing the U S government and public in the 1970s is whether the internationalist philosopy that has guided past policies should be followed in the future This question is raised by demands for change from two important sources First, U S labor, which believes that foreign investment results in a loss of domestic employment

I received much assistance in the preparation of foreign rates of taxation from the International Tax Affairs Staff of the U S Treasury Department, headed by N Gorden Valuable comments on an earlier draft of this paper were made by G Kopitz, P Gray, H Shapiro, M Feinberg, G Haberler, and other participants at a seminar at the U S Treasury Department, by an unknown referee of this *Journal*, and by my colleague Edward Tower

[1] For a review of the issues surrounding the internationalist policies see Krause and Dam (1967) and Musgrave (1969) These books also provide references to the original contributions to the field

opportunities and reduces the growth of wages, demands that the U S government "curb the substantial outflow of American capital" (Goldfinger 1971, p 928) Second, some major foreign governments have decided that long-term capital inflows from the United States are undesirable because they lead to "takeover" of their industries[2] and induce unemployment, causing deficits on current account, or result in an unwanted increase in official dollar holdings

In a scholarly analysis prepared for the Joint Economic Committee of Congress, Musgrave (1972, pp 177–78) concluded "There are sufficient doubts about the effects of foreign investment on the U S economy to lead to the conclusion that the U S tax treatment of foreign investment income should be reviewed and reevaluated This applies especially to deferral, but consideration may also be given to limiting the present credit for foreign taxes to less than 100 per cent Such measures would not be incompatible with opposition to trade restrictions Indeed, they might be supportive of free trade policy "

The present paper has been stimulated by this controversy over the merits of international investment in general and the effects of taxation in particular It attempts to shed light on the issues by presenting some estimates of the after-tax rates of return yielded by some forms of U S asset holdings abroad during the period 1960–69 In a rational decision about future U S policies, presumably such a rate of return should carry important weight

This paper represents a pioneering effort, and its primary aim is to sort out some theoretical issues and provide a rough first estimate No other studies of this sort seem to exist for the United States, though the broader issues of the efficiency aspects of international taxation agreements have been studied theoretically (Kindleberger 1968, Musgrave 1969) and estimates of the subsidies as dollar flows have been prepared (Musgrave 1972) The overall welfare effects of foreign investment have received a great deal of attention, and a ready set of references to these can be found in Reddaway (1968) and Dunning (1970) Jenkins (1972, chap 4) estimates Canada's taxation gains from foreign-owned capital in Canada

By concentrating the present study on the effects of taxation on the returns from only some forms of U S asset holdings, we neglect many considerations which should be weighted in the final decision about the future of international tax agreements Thus, we disregard the effects of capital flows on the terms of trade, income distribution, externalities, the efficiency of employment policies, and the benefits from the diversification of assets[3] Furthermore, the study does not consider the returns from such U S assets as bank loans, investment in oil and mining, and the tax and other benefits from foreign asset holdings in the United States.

---

[2] Significant political influence along these lines has long been felt in Canada and France Leftist factions in developing countries have a tradition of considering U S investments evidence of U S imperialism

In Part I current regulations concerning foreign investments are discussed briefly Parts II and III contain estimates of returns from foreign bond and equity holdings and direct investment in manufacturing, respectively The paper closes with a brief summary and conclusions

I. The Taxation of Foreign Income

The major feature of the U S taxation of income from assets held abroad is the tax offset provision of international taxation agreements in existence with all major Western countries of the world These tax provisions permit a U S taxpayer upon repatriation of his earnings to deduct from his tax liabilities to the U S government all taxes paid to foreign governments Assuming for a moment that all earnings are repatriated during the period in which they accrue, the U S tax rate on foreign investment income, t_d', is $t_d - t_f$, where t_d and t_f are the domestic tax rate on U S income and the foreign tax rate, respectively In cases where the foreign exceeds the domestic rate and therefore t_d' becomes negative, the taxpayer receives no explicit U S tax refund on a return from that country and $t_d' = 0$ whenever $t_d \gtreqless t_f$ But because the taxpayer has the option of combining returns from several countries, the overpayments in some countries can be used effectively to reduce obligations for U S tax payment on income from other countries, and in practice a refund is provided for overpayment to some foreign governments

As a result of the tax provision, the net private rate of return from foreign rather than domestic investment (R_p) is

$$R_p = (r_f)(1 - t_f - t_d') + r_x - r_d(1 - t_d), \tag{1}$$

where r_f and r_d are the foreign and domestic rates of return from foreign investment, respectively, and r_x is the nontaxable capital value change due to exchange rate revaluations expressed as an annual rate of return Under the assumptions that $t_f \leq t_d$, implying $t_f + t_d' = t_d$, and that $r_x = 0$, the net private after-tax rate of return from foreign investment is positive ($R_p > 0$) whenever

$$r_f > r_d \tag{2}$$

and it becomes zero when $r_f = r_d$ These conditions are well known to be a necessary requirement for the efficient international allocation of capital, though it should be noted that, because of the offset provisions mentioned in the preceding paragraph, if $t_f \geqq t_d$ and there are no tax credits from other countries, it is possible that $R_p < 0$ even though $r_f > r_d$ and $r_x = 0$ Therefore, the U S tax laws need not provide the necessary condition for the efficient international allocation of capital

---

[3] The classical theoretical article is MacDougall (1960) The gains from international diversification have been discussed in Grubel (1968) and Grubel and Fadner (1971)

Let us now define the social rate of return from foreign investment (R_s) as

$$R_S = (r_f)(1 - t_f) + r_x - r_d + g_d, \tag{3}$$

where g_d is the government expenditure necessitated by domestic investment but not charged to the business firm directly, expressed as an annual rate of expenditure per \$100 investment Under the assumption that r_x and g_d are zero, the social rate of return is positive ($R_S > 0$) only if

$$(r_f - r_d)/r_f > t_f, \tag{4}$$

which means that foreign investment is socially profitable only if the foreign rate of return exceeds the domestic rate by a percentage larger than the foreign tax rate The analysis also implies that foreign investment is privately profitable but results in a socially negative rate of return whenever $r_d/(1 - t_f) > r_f > r_d$

The task of the next parts of this paper is to estimate R_p and R_S using the variables shown in the right-hand side of equations (1) and (3) However, it will be impossible to provide any estimates of g_d, the savings of U S government expenditure realized by the location of the investment abroad, and g_d will be assumed to be zero It is difficult to know reliably the magnitude of the bias introduced by this assumption However, it seems clear that services provided free by governments specifically to private investment projects are a relatively minor proportion of all government expenditures, because most government services such as public utilities and judicial services are provided for a fee At the same time, such free services as roads and public education to the work force would most often be the same with and without the existence of the investments Final judgment of these matters must await more detailed theoretical analysis and some empirical case studies

The preceding analysis was set out by assuming that all earnings are repatriated in the period of accrual, and in the remainder of this paper we will retain this assumption We thus disregard the important fact that under present tax laws, whenever the foreign is below the U S tax rate, the payment of the remaining U S tax can be deferred until the profits are repatriated The owners of foreign-reinvested profits thus enjoy the benefits of a tax-free loan Musgrave (1972, p 192) has estimated that the revenue loss to the U S Treasury from this provision came to about \$900 million in the year 1970 for the case of foreign-incorporated subsidiaries of U S corporations

The arguments over the merits of this deferral provision exceed the scope of the present paper The primary economic impact of the implicit subsidy is an inefficiently high level of foreign investment, the welfare implications of which take the form of consumer surplus [4]

II. The Data and Rates of Return on Bonds and Equities

The empirical study of private and social rates of return to U S holdings and holdings abroad is based on average rates of return during the period 1960–69 This decade was characterized by the greatest degree of freedom from government restrictions in foreign exchange markets since the end of the Second World War and by rapid economic growth and capital formation throughout the world Of course, the results of this study would have been different if any other period had been analyzed This fact does not invalidate the calculations made, but limits the inferences we can draw about private and social profitability of foreign asset holdings over longer past periods and about the future Dunning (1970, p 83) has pointed to a difficulty associated with considering a period of rapid growth in foreign investments Since it is known that investments tend to yield a lower rate of return initially than over the longer run, average rates of return during periods of rapid growth are biased downward We have no way of quantifying this bias and will disregard it in the subsequent discussion

Overview of Total U S Foreign Investment Position

Before we turn to a detailed analysis of rates of return, it is useful to present an overview of the total U S foreign investment position in long-term assets in order to put into the proper perspective the asset categories to which the following estimates had to be limited For this purpose table 1 has been compiled The upper half of the table, showing U S assets, reveals in column 4 that the overwhelming proportion of foreign assets held abroad is in the form of direct investment (73 7 percent) followed by bonds (12 2 percent), corporate stocks (7 2 percent), and bank loans and other assets (6 9 percent) In column 3 we see that net growth during the decade 1960–69 was greatest in direct investments (122 1 percent) and in overall investments (116 1 percent)

In the last row of the upper half of table 1 we show a rough regional breakdown of U S assets Canada is the single most important location of U S -held foreign assets, with 33 9 percent of the total, followed by Western Europe (27 8 percent) and the rest of the world (38 3 percent) It is interesting to note that Canadian bonds and stocks represent 59 and 49

---

[4] For an excellent recent study of the welfare cost of U S trade restrictions using this approach and providing the relevant references to the literature, see Magee (1972)

TABLE 1

International Private Long-Term Investment Position of the United States ($ Millions)

	Total		Total Percentage		Canada		Western Europe		All Other	
	1960 (1)	1969 (2)	1960–69 (3)	1969 (4)	1969 (5)	% of All Countries (6)	1969 (7)	% of All Countries (8)	1969 (9)	% of All Countries (10)
					Assets					
Direct investments	31,865	70,763	122 1	73 7	21,075	29 8	21,554	30 5	28 134	39 7
Foreign bonds	5,524	11,712	112 0	12 2	6,912	59 0	607	5 2	4,193	35 8
Foreign corporation stocks	3,984	6,953	74 5	7 2	3,406	49 0	2,816	40 5	731	10 5
Bank loans and other	3,074	6,601	114 7	6 9	1,207	18 3	1,744	26 4	3,650	55 3
Total	44,447	96,029	116 1	100 0	32,600	33 9	26,721	27 8	36,708	38 3
					Liabilities					
Direct investments	6,910	11,818	71 0	28 8	2,834	24 0	2,659	22 5	6,325	53 5
Bonds	649	4,800	640 0	11 7	87	1 8	3,770	78 5	943	19 7
Corporate stocks	9,302	18,140	95 0	44 3	2,950	16 3	12,106	66 7	3,084	17 0
Banks and other	1,557	6,228	300 0	15 2	178	2 9	3,066	49 2	2,984	48 9
Total	18,418	40,986	122 5	100 0	6,049	14 8	21,601	52 7	13,336	32 5
Ratio of assets to liabilities	2 41	2 34			5 39		1 23		2 75	

Source —U S Department of Commerce, *Survey of Current Business*, various issues percent, respectively, of the U S holdings in these classes European corporate stocks also represent a significant 40 5 percent of the total The rest of the world has issued a disproportionately large amount of liabilities to U S banks (55.3 percent of the total U S bank loans) About one-third of these bank loans were to Japan

The bottom half of table 1 shows the level, composition, and growth of foreign holdings of U S assets The most important conclusions from this set of data are that overall U S assets were 2 41 and 2 34 times the level of liabilities in 1960 and 1969, respectively However, this average hides regional divergence, which produces a 1969 ratio of 5 39 for Canada, 1 23 for Europe, and 2 75 for the rest of the world Of further interest is the fact that for all countries U S corporate stocks represent the favorite form of asset, 44 3 percent, followed by direct investment, 28 8 percent This relationship is dominated by an overwhelming European preference of about 3 1 of stocks over direct investment Canadians hold roughly equal shares of U S direct investment and corporate stocks

The following empirical analysis is limited to the estimation of rates of return from U S holdings of foreign assets in the form of bonds, corporation stocks, and direct investments This partial approach has the serious shortcoming that the results do not provide information about the social profitability to the United States of adhering to the principle of not restricting international capital flows and partaking in a worldwide system of double taxation agreements In order to provide an estimate of this broad social profitability, it would be necessary to consider the benefits and costs arising from all forms of foreign assets and liabilities, long and short term, private and public, just as it is necessary to study all sources of income and costs to estimate the profitability of a private bank or business

This problem of partial estimates is particularly important, since the U S government tax revenues lost from U S holdings abroad are partly offset by the U S taxation of foreign investment income For these reasons, the present study must not be interpreted as providing an answer to the broad problem of U S benefits and costs from adhering to the internationalist principle in foreign investment policies The empirical estimates to be presented represent merely an input, though an important input, in such a broader calculus

Portfolio Investments

Portfolio investments are assets which do not confer control over the operation of the issuing business enterprise Qualifying as such are bonds and equities representing no more than 10 percent of the corporation's outstanding stocks

Calculations on the rates of return to bonds and stocks were limited to Canada and Western Europe, the major issuers of U S -held assets Data on U S foreign portfolio holdings do not disaggregate "Western Europe" by individual countries Average Western European rates of return were calculated by weighting individual country data by the value of U S manufacturing investment at the end of 1969, for which country data are available

Bonds

The U S government does not possess information on the actual composition or yield of U S private bond holdings abroad For this reason it was necessary to assume that the holdings consisted entirely of foreign government bonds with a maturity over 5 years and that the opportunity cost of these holdings was equal to the long-term yield on U S government bonds These assumptions produce unbiased results as long as (1) the yield differentials between government and private bonds are equal in the United States and the rest of the world, (2) U S wealth holders purchase equal proportions of government and private bonds abroad as they would if they had invested domestically, (3) the maturity composition of foreign portfolios is the same as the domestic portfolio would have been and there are no international differences in yield curves, and (4) income from foreign government bonds is taxable at the normal rate [5] Refinements of the estimates will have to await information on the accuracy of these hypothetical propositions about the real world

Row 1 of table 2 shows the foreign bond yields (r_f) to have been 5 89 and 6 28 percent for Canada and Western Europe, respectively Comparison of these yields with the U S yield (r_d) shown in row 8 as 4 51 percent reveals a positive net private profitability from foreign bond holdings The foreign withholding taxes (t_f) of 15 0 and 3 2 percent (row 2) lower after-tax foreign yields to 5 01 and 6 08 percent (row 3) for Canada and Western Europe, respectively Average exchange rate losses during the period (r_x) shown in row 4 and applied to the proportion of non-dollar-denominated bonds (row 5) reduce the gross foreign yields to 4 92 and 6 06 percent, respectively, for Canada and Western Europe (row 7) After adjustment for the domestic opportunity cost of these funds, the net social rate of return (R_s) is estimated as having been 0 41 and 1 55 percent for Canada and Western Europe, respectively

Three brief comments regarding these calculations should be made

---

[5] Canadian bonds issued by the provinces and municipalities especially tend to be free from Canadian taxation This fact biases downward our estimate of U S rates of returns from bond holdings Given all of the other simplifying assumptions, in the extreme, if all Canadian bonds held had been tax free, the net social rate of return would have been 1 29 rather than 0 41 percent

TABLE 2

U S RETURNS FROM FOREIGN BOND HOLDINGS, 1960–69

	Canada	Western Europe
1 Average yield, government bonds, 1960–69, %/year	5 89	6 28
2 Average withholding tax, 1962–67, % of interest	15 0	3 2
3 After-tax rate of return, %/year	5 01	6 08
4 Average exchange rate change, 1960–69, %/year	−0 74	−0 46
5 Percentage of U S -dollar-denominated bonds 1969	88 4	96 0
6 Change of portfolio due to exchange rate	−0 09	0 02
7 After-tax, after-exchange rate adjustments, return, %/year	4 92	6 06
8 Average U S government bond yield, 1960–69, %/year	4 51	
9 Net social rate of return, %/year	0 41	1 55

SOURCES AND NOTES ON COMPUTATIONS

Yield Organization for Economic Cooperation and Development (1970, vol 2) Fourth-quarter average long-term government bond yields were averaged over 1960–69 for each country of Western Europe Weights for calculation of Western European average are relative value of U S manufacturing investment in 1969 from U S Department of Commerce, *Survey of Current Business* (October 1970)

Tax rate Statutory dividend withholding tax rates compiled by U S Treasury International Tax Affairs Staff (Kopits 1970) Average for Western Europe weighted by U S manufacturing investment

Exchange rate change End-of-period exchange rates 1960 69 from International Monetary Fund (various years) Change in rate expressed as percentage of 1960 rate European rate is average weighted by manufacturing investment

U S bond rates 3–5-year U S government issues (U S Government 1971, p 264 table C57)

First, the exchange rate adjustment would be much more unfavorable if the December 1971 currency realignments had been within the period covered in this study However, the net effect on the foreign rates of return would not have been mitigated considerably by the fact that nearly 90 percent of all foreign bonds held by U S residents are denominated in dollars

Second, the U S bond yield shown is unadjusted for the theoretical fall in the U S rate that would have occurred if the funds invested abroad had been used to purchase U S bonds This adjustment seems unnecessary in light of the fact that the total value of bonds in circulation in the United States at the end of 1969 was $495 billion, of which the $7 5 billion of foreign bonds represent only 1 5 percent

Third, it is customary to assume that market interest rates reflect expected rates of inflation over the life of financial assets The risks of unexpected changes in the price level and the resultant wealth redistribution are shared equally between lenders and borrowers, assuming that expectations on the average are correct However, when unexpected inflation does occur and is biased in favor of foreign borrowers, the resultant wealth gain accrues to non-U S residents [6] One may wonder

---

[6] Aliber (1973) has shown that during several years of the post-World War II period interest rates typically failed to reflect exchange rate changes

TABLE 3

U S RETURNS FROM FOREIGN CORPORATION STOCK HOLDINGS, 1960–69

<table>
<tr><th></th><th></th><th>Canada</th><th>Western Europe</th></tr>
<tr><td>1</td><td>Average dividend yield, 1960–69, %/year</td><td>3 4</td><td>3 9</td></tr>
<tr><td>2</td><td>Average capital gain, 1960–69, %/year</td><td>9 6</td><td>2 7</td></tr>
<tr><td>3</td><td>Total yield (1) + (2)</td><td>13 0</td><td>6 6</td></tr>
<tr><td>4</td><td>Average withholding tax, % on dividends</td><td>15 0</td><td>7 6</td></tr>
<tr><td>5</td><td>Average capital gains tax, %</td><td>0 0</td><td>11 3</td></tr>
<tr><td>6</td><td>Net after-tax yield, %/year</td><td>12 5</td><td>6 0</td></tr>
<tr><td>7</td><td>Average exchange rate change, 1960–69, %/year</td><td>−0 74</td><td>−0 46</td></tr>
<tr><td>8</td><td>Yield after tax and exchange loss</td><td>11 7</td><td>5 5</td></tr>
<tr><td>9</td><td>Average U S dividend yield, 1960–69, %/year</td><td colspan="2">3 2</td></tr>
<tr><td>10</td><td>Average U S capital gain, 1960–69, %/year</td><td colspan="2">6 7</td></tr>
<tr><td>11</td><td>Sum (9) + (10)</td><td colspan="2">9 9</td></tr>
<tr><td>12</td><td>Net social rate of return (8) − (11)</td><td>1 8</td><td>−4 4</td></tr>
</table>

SOURCES AND NOTES ON COMPUTATIONS

Dividend yield Annual yield for Canada, United Kingdom France, Germany Italy, the Netherlands Norway and Sweden from Organization for Economic Cooperation and Development (1970), each country's annual rate is averaged for the decade Western Europe average weighted by book value of U S direct investment in manufacturing in 1969 (from U S Department of Commerce *Survey of Current Business*, October 1970)

Capital gains Based on stock market indices of all countries in sample Source same as dividend yield Western Europe average weighted like dividend yield figure

Withholding tax rate and exchange rate changes see table 2

U S dividend and capital gains Standard and Poor's 500 stocks (Organization for Economic Cooperation and Development 1970 pp 162 65) Last month of each year averaged over 10 years for dividend yield Percentage gain of price end of 1960 to end of 1969 for capital gains

to what extent the average nominal bond yields shown above can be assumed to have reflected fully anticipated rates of inflation during the period [7] However, it is analytically not useful to consider that investors consistently over a decade are injured by unexpected inflation rates, and the data presented above will not be corrected for any inflation losses

Equities

The estimation of foreign rates of return on equities is nearly identical to that undertaken for bonds In the absence of any information on yields of actual holdings, the yield figures of rows 1–3 are based on the assumption that U S holders owned portfolios equal to the "averages" of the individual markets, with Western European holdings weighted by the share of U S direct investment As can be seen from table 3, private rates of return from Canadian stocks were equal to 3 4 and 9 6 percent from dividends and capital gains, respectively The analogous rates for Europe were 3 9 and 2 7 percent After application of the withholding and capital gains taxes (rows 4 and 5), the after-tax yields were 12 5 and 6 0 percent for Canada and Europe, respectively Adjustments for exchange rate losses yielded the final figures of 11 7 and 5 5 percent for the two regions

---

[7] During the period 1960–69, when the U S consumer price index rose 23 8 percent, analogous percentage rises in some of the major countries in this study were Canada, 26 7, United Kingdom, 40 0, Germany, 25 7, France, 42 5, Italy, 40 3 (International Monetary Fund, various years, country statistics)

One theoretically very difficult problem I have been unable to solve concerns the choice of an appropriate U S opportunity cost for the funds invested in foreign equities There appear to be two extreme theoretical possibilities On the one hand, we may consider that the purchase of domestic rather than foreign equities would have affected only the stock of U S financial wealth in the form of equities in the portfolios of the public, there would have been no induced new equity flotations, corresponding real capital formation, and profits taxable at the corporate income tax rate Under these assumptions and given that foreign equity holdings are but a minute fraction of the stock of U S shares held by Americans, the proper opportunity cost of foreign equity purchases is the returns from U S stock market investment Rows 9–11 in table 3 show the dividend yield and capital gains of stocks in Standard and Poor's 500 share index They total 9 9 percent and imply that the holdings of Canadian shares resulted in a positive net social rate of return of 1 8 percent per year during the decade The corresponding figure for Western Europe was a negative 4 4 percent

The other extreme treatment of the opportunity cost of foreign equity purchases considers them to have come fully at the expense of real capital formation by existing or potential new firms In order to produce after-tax profits for dividend distribution and reinvestment equal to the 9 9 percent annual yield on equity shares, these firms would have had to earn a pretax rate of return of 16 6 percent, assuming the corporate tax rate to be 40 4 percent (see table 4, row 4 and note *h*) While the preceding arguments contain many hidden assumptions about the operation and efficiency of capital markets and the role of taxes and depreciation, the resultant estimate of the rate of return on real capital formation is close enough to other data available (see table 4, row 3, last column, and note *h* for the source) to provide a reasonably reliable benchmark for our purposes of calculation Use of the preceding assumptions leads to the conclusion that the average annual rates of loss from U S equity holdings during the decade were 4 9 and 11 1 percent for Canada and Europe, respectively

Without much further empirical and theoretical analysis, which would lie outside the main objectives of this paper, we cannot decide which of the above two theoretical models describes the real world most accurately However, the ranges calculated under the two extreme assumptions indicate that foreign equity investments during the period probably yielded a negative rate of return, though a very small positive yield from Canadian equity holdings cannot be ruled out

TABLE 4

AVERAGE RATES OF RETURN FROM DIRECT MANUFACTURING INVESTMENT, 1960–69

	Canada	U K	Belgium and Luxembourg	France	Germany	Italy	Netherlands	Denmark
1 Book value, 1969, $ million	9,389	4,555	700	1,518	2,750	716	656	58
Before-tax rates of return on								
2 Book value	16.4	21.6	23.5	14.3	26.7	17.0	15.3	20.1
3 Equity	22.6	28.1	30.1	18.6	34.3	22.1	19.9	26.1
4 Effective corporate tax rates	45.0	45.4	31.2	37.0	33.0	44.2	24.4	13.9
After-corporate-tax rates of return on								
5 Book value	9.0	11.8	16.2	7.6	17.8	9.5	11.6	17.3
6 Equity	12.4	15.3	21.1	9.9	23.1	12.4	15.1	22.5
7 Ratio of dividends/earnings	0.48	0.56	0.37	0.46	0.60	0.72	0.63	0.97
8 Tax rate on dividends	15.0	0.0	15.0	15.0	15.0	5.0	0.0	5.0
9 After corporate and dividend taxes rate of return, equity	10.8	15.3	19.4	8.6	20.0	11.6	15.1	21.2
10 Exchange rate adjustment, % per year	−0.7	−1.4	0.0	−1.4	1.2	0.0	0.4	−0.8
11 Gross social rate of return, (8) − (9)	10.1	13.9	19.4	7.2	21.2	11.6	15.5	20.4
12 Net social rate of return, (10) − 19.3%	−9.2	−5.4	0.1	−12.1	1.9	−8.3	−3.8	1.1
13 After adjustment for royalties and fees	−6.3	−0.3	5.2	−7.0	7.0	−3.2	1.3	6.2

Table 4 (*Continued*)

	Norway	Spain	Sweden	Switzerland	Australia	South Africa	Sum or Average	U.S. Domestic
1 Book value, 1969, $ million	62	295	179	380	1,567	374	23,199	
Before-tax rates of return on								
2 Book value	15.6	12.9	23.3	21.4	23.7	27.2	19.5	
3 Equity	20.3	16.8	30.3	27.8	30.8	35.4	25.5	19.3
4 Effective corporate tax rates	32.1	30.3	48.5	15.8	39.3	25.8	40.7	40.4
After-corporate-tax rates of return on								
5 Book value	10.6	9.0	12.0	18.0	14.4	20.2	11.5	
6. Equity	13.8	11.7	15.6	23.4	18.7	26.3	15.1	11.5
7 Ratio of dividends/earnings	0.00	0.66	0.92	0.46	0.49	0.55	0.52	
8 Tax rate on dividends	5.0	15.0	5.0	5.0	15.0	7.5	10.9	
9 After corporate and dividend taxes rate of return, equity	13.8	10.0	14.2	22.8	16.4	24.8	13.7	
10 Exchange rate adjustment, % per year	0.0	−1.7	0.0	0.0	0.0	0.0	−0.3	
11 Gross social rate of return, (8) − (9)	13.8	8.3	14.2	22.8	16.4	24.8	13.4	
12 Net social rate of return, (10) − 19.3%	−5.5	−11.0	−5.1	3.5	−2.9	5.5	−5.9	
13 After adjustment for royalties and fees	−0.4	−5.9	0.0	8.6	2.2	10.6	−0.9	

SOURCES AND NOTES ON COMPUTATIONS

a) Book value from U.S. Department of Commerce *Survey of Current Business* (October 1970), p. 23

b) After-tax rates of return on book value are equal to "earnings" divided by preceding year's end "book value" (U.S. Department of Commerce, *Survey of Current Business*, various years)

c) Equity is equal to book value divided by book value minus liabilities to parent (from U.S. Department of Commerce, *Foreign Affiliate Financial Survey*, 1966–69 [1970, pp. 23–33, tables III and IIIb])

d) Effective tax rates from U.S. Treasury Department (1969, p. 143, table 24)

e) Tax rate on dividends supplied by U.S. Treasury Department

f) Exchange rates from International Financial Statistics (various years)

g) Dividends are equal to "income" from direct foreign investment (U.S. Department of Commerce, *Survey of Current Business* various issues)

h) U.S. domestic rates of return on manufacturing investment from U.S. Government (1969 p. 283, table C74). Tax rate calculated from given rates

i) Royalties and fees from U.S. Department of Commerce *Survey of Current Business* various issues

All rates of return for individual countries are simple average of period 1960–69. World average is weighted by 1969 value of direct investment

III. Direct Investment in Manufacturing

United States statistics on direct foreign investment are much more complete and reliable than the statistics on stocks and bonds For this reason, in row 1 of table 4 we show U S holdings by individual countries, the sum of which represents 78 8 percent of all U S direct manufacturing investment abroad in 1969

United States direct investment holdings at the end of 1969 were 41 6 percent in manufacturing, 8 0 percent in mining and smelting, 28 2 percent in petroleum, and 22 2 percent in other industries The present study is limited to manufacturing investment for two reasons First, because the profitability of investment in mining and petroleum is so dependent on the availability of natural deposits, it is impossible to establish reliably what the domestic opportunity cost of the foreign investment would have been Second, special treaties with foreign governments about royalties, special accounting, and depreciation rules governing these nonmanufacturing industries abroad and in the United States makes it impossible to estimate rates of return which would be theoretically appropriate for the present purposes of analysis

The U S government publishes regularly "earnings" and the "book value" of U S manufacturing enterprises abroad Earnings are defined as "the sum of the U S share in net earnings of foreign corporations and branch profits after foreign but before U S taxes " Book value is defined as "equity of, plus liabilities to, parent "[8] The published U S statistics on rates of return from foreign investment simply represent earnings per year as a percentage of book value at the beginning of that year A 10-year average of these rates of return (1960–69) is shown in row 5 of table 4 In the following we will discuss the adjustment made on these figures to derive the estimates of net social rates of return shown in rows 12 and 13 of table 4

Earnings and Foreign Income Taxes

In order to compare properly U S and foreign business income tax rates, it is insufficient to use statutory rates because in each country definitions of income, expenses, depreciation allowances, etc , are different Fortunately, there exists a U S government study of effective foreign tax rates based on the U S definition of corporate income and depreciation rates (U S Treasury Department 1969) These tax rates were estimated from data submitted to the Internal Revenue Service for the fiscal year 1961–62 by calculating for each corporation its income according to U S laws and expressing the actual income taxes paid to foreign governments as a fraction of this income As can be seen from row 4 of table 4, these tax rates vary widely between countries, but the weighted average of 40 7 percent is only 0 3 percent above the domestic U S rate

---

[8] For an updated and clear definition of terms employed in U S government statistics on the U S foreign investment position see U S Department of Commerce, Survey of Current Business (October 1971, pp 37–38)

As a result of the absence of foreign tax rate data on years other than 1961–62, it was assumed that these rates were in force throughout the decade under study, and the average before-tax rates of return for the period were calculated by inflating the average after-tax rates of row 5 The results are shown in row 2 of table 4 The weighted average of before-tax returns on book value is 19 5 percent per year

Book Value versus Equity

However, before domestic and foreign rates can be compared, it is necessary to adjust the foreign rates so that they reflect return to equity rather than book value The liabilities to parent, which represent the difference between foreign equity and book value, are loans or other extensions of credit, which do not share in the risks accruing to equity investment The earnings on these nonequity U S assets should be considered as a separate category of U S foreign assets but will be disregarded here In this study we will adjust the book value by taking advantage of a recent publication (U S Department of Commerce 1971) of the financial structure of U S -owned corporate manufacturing affiliates abroad It shows liabilities to parents to have been 27 3 and 23 1 percent of book value for affiliates in Canada and the rest of the world, respectively, during the period 1966 69 Under the assumption that this financial structure is representative of the full population of U S -owned foreign corporations and branches[9] and for the full decade 1960 69, we estimated average annual rates of return to equity by multiplying the rates of return on book value shown in rows 2 and 5 by 1 376 and 1 30 for Canada and the remaining countries, respectively The results are shown in rows 3 and 6 of table 4 In the remaining parts of table 4 all further adjustments on rates of return will be made on the equity base, which is the theoretically proper one for the comparisons with U S domestic rates of return

Dividend Withholding Taxes

United States income from foreign manufacturing investment is subject to double taxation in the same way as income from domestic corporations First, the government taxes the corporation's income directly, which calls for the adjustments already made Then the dividends distributed by the corporation are subject to a dividend withholding tax In row 8 of table 4 we show the rate at which foreign governments withhold dividends

However, the second form of taxation takes place only on that portion of earnings that is distributed rather than reinvested In row 7 of table 4 the average 1960–69 ratio of distributed to before-income-tax earnings is shown For all countries (next to last column of table 4) the average of 10 9 percent withholding tax (row 8) applied to the 52 percent average distributed earnings (row 7) is seen to lower the after-corporate-income-tax rate of return on equity of 15 1 percent (row 6) to 13 7 percent (row 9) Therefore, the average withholding tax rate is effectively equal to 1 4 percent of equity per year

---

[9] According to U S Department of Commerce, *Survey of Current Business* (October 1971, p 38), earnings of foreign corporations and foreign branches were $6 3 and $2 5 billion, respectively, in 1970

Exchange rate changes over the period under study increased or decreased the capital value of the foreign investment holding in terms of U S dollars at the annual rate shown in row 10 of table 4 Row 11 contains the rates of return after taxes and exchange rate adjustments, which are the theoretically relevant base of comparison with the domestic opportunity cost of the resources, the pretax rates of return on U S domestic manufacturing investment of 19 3 percent during the decade (see note *h* of table 4 for the source of the 19 3 figure) Row 12 of table 4 presents the net social rates of return of investing abroad rather than domestically as defined in equation (3)

As can be seen, the all-country average shows a negative yield of −5 9 percent per year Positive yield of 0 1, 1 9, 1 1, 3 5, and 5 5 percent were achieved only from the investments in Belgium and Luxembourg, Germany, Denmark, Switzerland and South Africa, respectively

Royalties and Fees

The theoretical literature on the determinants and welfare effects of direct foreign investment[10] stresses the comparative advantages accruing to multinational corporations from the use of research, management, and marketing skills, which are developed at high initial cost for the domestic U S market but which can be applied at relatively low marginal private cost in the operation of foreign subsidiaries In their accounting conventions, these corporations charge their subsidiaries the average cost of developing the production, management, and marketing knowledge, and the U S government reports the value of income from these sources as "royalties, license fees, and rentals" and "management fees and charges," royalties and fees for short

In the reporting of the balance of payments it is theoretically unobjectionable to include royalties and fees in "income from foreign investment," along with dividends and distributed profits,[11] on the grounds that without the foreign direct investment this financial flow would not have taken place However, in the present analysis we are concerned with the productivity of resources, not financial flows From this point of view the full value of royalties and fees represents a net increase in the productivity of U.S knowledge capital only if the marginal cost of providing these services is genuinely zero More generally, the social income from royalties and fees charged foreign subsidiaries is equal to the difference between the average cost at which they are billed and the true marginal cost

---

[10] See Kindleberger (1968), Johnson (1969), and Caves (1971)

[11] The U S balance-of-payments statistics report income from foreign investment as the sum of dividends and distributed profits plus royalties and fees to subsidiaries

Unfortunately, there exists no evidence on the relationship between marginal and average cost of these services, and in our calculations we had to assume that the marginal cost was zero [12] As a result, the estimate of the return from foreign investment is almost certainly biased upward, since it is highly unlikely that the marginal cost is zero However, as one of the important findings of this paper is the negativity of the net social rates of return from foreign investment, the bias introduced by our assumption is in the proper direction

During the period 1960–69, the value of royalties and fees averaged 2 1 and 3 9 percent annually of the book value of foreign investment in Canada and Europe, respectively These figures were adjusted for the equity base, and we assumed that Australian and South African rates were the same as those for Europe The resultant increase in rates of return, which is not subject to any forms of foreign taxation, was added to the net social rates of return shown in row 12 of table 4 The results are presented in row 13 of table 4 and indicate that the overall, all-country average was −2 7 percent compared with −6 4 percent without the adjustment for royalties and fees Positive adjusted rates of return are shown for Belgium and Luxenbourg (5 2), Germany (7 0), the Netherlands (1 3), Denmark (6 2), Switzerland (8 6), Australia (2 2), and South Africa (10 6) The overall average (−0 9) is dominated by the proportionately large investments in Canada (−6 3) and the United Kingdom (−0 3) France (−7 0) and Spain (−5 9) retained their substantial negative rates

Confiscation

In the past, U S investments in many countries of the world, but especially in Cuba, Eastern Europe, China, and some Latin American countries, have been lost to private investors and to the U S economy as a result of confiscation or inadequate compensation by foreign governments In principle, these losses should be averaged and incorporated into the present calculation of net social rates of return from foreign investment Unfortunately, no ready data are available to carry out such calculations, and the necessary adjustments of our estimates are not undertaken Consequently, the results presented above are biased upward In spite of this upward bias, however, the rates still average out to negative

---

[12] Theoretically, the opportunity to earn further income on the knowledge capital originally produced only for domestic use should increase the yield from this knowledge capital relative to other forms of capital formation Therefore, the opportunity to charge foreign subsidiaries for a share of these costs should induce increased expenditures Once such an expansion has taken place, it is theoretically appropriate to charge foreign subsidiaries the full share of the average cost and inappropriate to consider income from royalties and fees as representing a net social gain acquired at zero or even very low marginal cost In the long run all fixed costs are variable, and the analysis should reflect this fact For this reason the adjustments to the rates of return on row 12 must be considered as introducing a theoretically very strong upward bias to the results

IV. Summary and Conclusions

The analysis presented in this paper must be considered exploratory Future studies may well lead to important revisions of the main conclusions that during the decade 1960–69 U S holdings of foreign bonds were socially profitable while some of the equity and direct manufacturing holdings resulted in a social loss Extensions of the analysis in future studies may profitably take place along the following lines

First, an attempt should be made to estimate returns to total U S asset holdings abroad, including short-term assets, bank loans, direct investment in forms other than manufacturing, government holdings, etc The effect of investment in mining and oil on U S terms of trade should be analyzed carefully

Second, the U S holdings should be considered as part of a global system which results in foreigners also holding U S assets taxable by the U S government For example, at the end of 1969 the nonliquid, private, long-term U S investment position showed $96 billion assets abroad and $41 billion liabilities to foreigners Foreign investment in the United States probably results in net U S gains which can be considered as offsetting the losses discussed above

Third, the externalities and government expenditures associated with asset holdings of various forms should be explored further As an entity, U S taxpayers receive government services that tend to be equal in value to all of the taxes paid The relevant question about the welfare effects of foreign direct investment therefore is the extent to which tax payments by corporate or other business entities can be considered as being in return for a government service that can be eliminated in the United States and has to be undertaken abroad because of the U S capital export

Until these lines of analysis have been explored further, the results of the present paper on the social nonprofitability of many foreign investments should be used with the proper caution in any policy decisions about restrictions on and taxation of U S asset holdings abroad However, the evidence is sufficiently reliable to conclude that the export of capital from the United States for direct investment in manufacturing and the internationalist policies guiding the taxation of income from this capital do not clearly raise U S welfare and probably lower it

References

Aliber, R "Tests of Fisher's Expectations Hypothesis " Paper presented at the Royal Economic Society Conference of Specialists in International Monetary Economics, Ditchley Park, England, January 1973

Caves, R "International Corporations The Industrial Economics of Foreign Investment " *Economica* 38 (February 1971) 1–27

Dunning, J *Studies in International Investment* London George Allen & Unwin, 1970

Goldfinger, N "A Labor View of Foreign Investment and Trade Issues " In *Williams Commission Report* Washington Government Printing Office, 1971

Grubel, H G "Internationally Diversified Portfolios Welfare Gains and Capital Flows " *A E R* 58 (December 1968) 1299–1314

Grubel, H G , and Fadner, K "The Interdependence of International Equity Markets " *J Finance* 26 (March 1971) 89–94

International Monetary Fund *International Financial Statistics* Washington Internat Monetary Fund, various years

Jenkins, G "Analysis of Rates of Return from Capital in Canada " Ph D dissertation, Univ Chicago, 1972

Johnson, H G "The Efficiency and Welfare Implications of the International Corporation " In *The International Corporation*, edited by C P Kindleberger Cambridge, Mass M I T Press, 1969

Kindleberger, C P *American Business Abroad* New Haven, Conn Yale Univ Press, 1968

Kopits, G F "The Industrial Economics of Foreign Investment A Comment " Mimeographed Washington Treasury Department, 1970.

Krause, L , and Dam, K *Federal Tax Treatment of Foreign Income* Washington Brookings Inst , 1967

MacDougall, D "The Benefits and Costs of Private Investment from Abroad " *Econ Record* 36, no 73 (March 1960) 13–36

Magee, S "The Welfare Effects of Restrictions on U S Trade " Working Paper no 18 Chicago Graduate School of Business, University of Chicago, December 1972

Musgrave, P *United States Taxation of Foreign Investment Income* Cambridge, Mass Harvard Law School, International Tax Program, 1969

——— "Tax Preferences to Foreign Investment " In *The Economics of Federal Subsidy Programs*, pt 2 Washington Joint Econ Committee of Congress, June 11, 1972

Organization for Economic Cooperation and Development *Financial Statistics*, vol 2 Paris Org Econ Cooperation and Development, December 1970

Reddaway, W B , et al *Effects of U K Direct Investment Overseas* Cambridge Cambridge Univ Press, interim report 1967, final report 1968

U S , Department of Commerce *Survey of Current Business*, various years Washington Dept Commerce

——— *Foreign Affiliate Financial Survey, 1966–69* Washington Dept Commerce, July 1971

U S , Government *The Economic Report of the President* Transmitted to Congress February 1971 Washington Government Printing Office, 1971

U S , Treasury Department *Internal Revenue Service, Statistics of Income, 1962, Supplemental Report Foreign Income and Taxes Reported on Corporation Income Tax Returns* Washington Government Printing Office, 1969

Cattle as Capital Goods and Ranchers as Portfolio Managers: An Application to the Argentine Cattle Sector

Lovell S Jarvis

University of California, Berkeley, and the Ford Foundation, Santiago, Chile

Microeconomic models, treating cattle as capital goods and producers as portfolio managers, provide a theoretical framework for an econometric model of the Argentine cattle sector Argentine agricultural stagnation has been partially attributed to a lack of producer price response, but this paper shows that the cattle sector exhibits strong price response and that producers correctly differentiate their behavior toward animals of different age and sex The long-run price response of slaughter is positive, but the short-run response is negative because animals must be withheld to permit increases in future output This negative short-run response implies that devaluation is unlikely to increase beef exports for at least 2 years

In this paper a number of microeconomic models are developed to provide a theoretical framework on which an econometric model of the Argentine cattle sector could be based The central theme is simple Cattle are considered to be capital goods which are held by producers as long as their capital value in production exceeds their slaughter value In essence, producers become portfolio managers seeking the optimal combination of different categories of animals to complement their non-cattle assets, given existing conditions and future expectations The models developed show clearly the differential impact of parameter changes upon the capital values of animals of different age and sex and indicate that the equations explaining slaughter and average slaughter

---

This paper is a revision of chapters 2 and 8 of my Ph D thesis, "Supply Response in the Cattle Sector The Argentine Case 1937/38–1966/67," Department of Economics, Massachusetts Institute of Technology, 1969 (which is to be published by the University of California Press) Many individuals provided helpful criticisms and encouragement at various stages of the study I would like to thank especially Franklin Fisher, Charles Freedman, Richard Schmalensee, J Phillip Cooper, Carlos Díaz-Alejandro, George Akerlof, and Steven Goldman Financial assistance was received from the National Science Foundation, the Center for Latin American Studies at the University of California, Berkeley, and the Ford Foundation The theoretical section of this paper was completed in 1968 while I was a guest of the Instituto Torcuato di Tella, Buenos Aires, Argentina

weight in an econometric model should be disaggregated by animal categories to obtain a meaningful explanation of producers' responses. Such an econometric model was developed and estimated for the Argentine cattle sector for the years 1937–67. The econometric model does a very good job, judging by conventional statistical tests, of explaining the operation and past behavior of the Argentine cattle sector, and the results comply closely with the previously developed microtheory.

There is considerable historical interest in the functioning of the Argentine cattle sector, long a prototype of large-scale traditional ranching in Latin America and the source of a significant proportion of world beef exports (nearly 20 percent during the 1960s). This interest has become more sharply focused in recent years because of the stagnation of Argentine agricultural production and the consequent desire in Argentina to formulate an improved governmental policy toward agriculture.[1] Among the factors alleged to be a cause of stagnation is a lack of response by agricultural producers to economic incentives, particularly to price changes. Several efforts were made during the 1960s to test empirically the price responsiveness of Argentine agricultural producers. These studies sought to at least partially answer the questions regarding producer behavior, and to provide policy makers with a framework for predicting the effect on production (and exports) of changes in agricultural prices. Somewhat surprisingly, given the experience in most other countries where similar studies have been carried out, little price response has been discovered in Argentina.[2] Almost no response was found for corn, a major crop which has experienced considerable output variation, and the response evident during the 1945–65 period for wheat, linseed, and oats is of much smaller magnitude and lower statistical significance than that encountered for these crops in most countries. This apparent lack of price response gave support to the "structuralist" interpretation of stagnation (and suggested that price policy, per se, may have very limited usefulness in Argentine agriculture).

---

[1] The traditional exports of Argentina are produced in the Pampa, one of the world's richest agricultural areas. These products (food grains, oil seeds, wool, and cattle) still accounted for 93 percent of total exports in 1965, despite the fact that gross agricultural output in this area showed almost no increase between 1935–39 and 1960–64. Thus, stagnation here implied stagnation in total exports, with serious consequences for Argentine development in general. Moreover, beef production retains the most important position among agricultural activities. Not only has beef accounted for more than 25 percent of total exports in recent years, but it also is one of the principal wage goods in Argentina, weighing heavily in the cost-of-living index and being a particularly important political symbol to the working classes.

[2] Diaz-Alejandro (1965, 1970), Colomé (1966), Freire (1966), Williams (1966), and Reca (1967), all estimated acreage or output response equations for different products, and all found little price response. Diaz-Alejandro, for example, concluded that other factors such as soil erosion, weather, the availability of inputs, social overhead facilities, and credit affected producers more than did prices and had, on balance, shifted the supply schedule to the left over time.

There is some previous evidence of price responsiveness for cattle production qualitatively consistent with the results reported here, but earlier studies usually did not encounter significant price coefficients.[3] The models used generally had a very simple specification, however, and the results obtained could have been attributable to the omission of important variables. These earlier studies were also plagued by a lack of reliable data, which made suspect their econometric results.[4] And the time profile of the response of beef supply to price changes also confused many observers.

The cattle sector presents an interesting feature insofar as the slaughter of animals responds negatively to a price increase in the short run. This behavior contrasts with the supply response of most other agricultural products, such as field crops, whose desired output (in the Nerlovian sense) and actual output are both expected to rise immediately in response to a price increase. Although actual output for these crops is expected to adjust only gradually to the long-run desired output, there is never reason to expect output to fall when its price increases.[5] But, because cattle production can be increased only by increasing the size of the breeding herd and/or withholding animals for further fattening, producers must bid animals away from consumers to increase the capital stock which is the source of higher future beef production. And the slow rate of biological reproduction causes the negative supply response to persist for some time. Because of the continuing confusion within Argentina regarding the cause and implications of the short-run reduction in slaughter, I sought to show that this behavior was rational, that a properly specified model would show a significant price response during the 1937–67 period, and that one could even show a theoretically correct differentiation of producers' behavior toward animals of different age and sex.[6]

I

To begin, micro-models are used to determine the optimum slaughter age and feed input for a steer, given growth functions for the animal and the following parameters faced by producers: the price of beef, the interest rate, and the cost of other inputs To simplify the exposition, I begin with a model where the only input is the steer itself[7] Let θ = age of the steer, $w(\theta)$ = weight of the steer at age θ, $\partial w/\partial\theta > \theta$, $\partial^2 w/\partial\theta^2 < 0$, r = interest rate, and $V(\theta)$ = the present discounted value of an animal allowed to live to age θ Thus if $V(\theta) = w(\theta)e^{-r\theta}$, where the price of beef is arbitrarily fixed at a constant value of unity, the steer will be slaughtered at age $\hat{\theta}$, which is chosen to maximize $V(\theta)$ The first-order condition for a maximum yields the requirement $(\partial w/\partial\theta)/w = r$, $\hat{\theta}$ occurs when the rate of growth of the animal is equal to the interest rate The second-order condition, $\partial^2 V/\partial\theta^2 < 0$, follows easily and requires that the rate of gain be declining In this model there is no opportunity cost to cattle production other than the interest foregone on invested capital An increase in r will lower the optimal slaughter age, and vice versa

---

[3] See Diaz-Alejandro (1965), Colome (1966), Otrera (1966), and Reca (1967).

[4] The required data series for herd stocks were constructed previously, making use of official censuses (as benchmarks) and official slaughter data. These constructed series were subjected to considerable independent testing for possible bias, a detailed discussion of which is contained in Jarvis (1969) and presented in revised form in Jarvis (1971), as an example of how economic models may be used to construct unavailable data from existing data. The reader should be aware that remaining biases, or those introduced in the construction process, could affect the results presented in this paper.

[5] Marketed output can decline if the increase in producer self-consumption exceeds absolutely the rise in output. See Behrman (1967).

[6] Diaz-Alejandro (1965) provided the first clear outline of the theory developed here, pointing out that a relative rise in beef price induces producers to invest further in beef production, which can be accomplished only be reducing current slaughter.

$$\frac{\partial\hat{\theta}}{\partial r} = -\frac{\partial^2 V/(\partial\theta\,\partial r)}{\partial^2 V/\partial\theta^2} < 0,$$

for (1)

$$\frac{\partial^2 V}{\partial\theta\,\partial r} = e^{-r\theta}\left(r\theta w - \frac{\theta\,\partial w}{\partial\theta} - w\right) < 0$$

Although this model implies an important role for the interest rate, the latter actually plays a relatively unimportant role, as is discussed subsequently

The model can be made more realistic by recognizing that a steer requires certain costly inputs throughout his life, which must be considered when choosing the optimal slaughter age Although the slaughter decision still depends upon the animal's rate of growth, the interest rate, and prices, it is conditional on the animal's being "fed" the optimal ration The criterion becomes maximization of the present discounted "profit" of the fattening process, which in perfect markets will be the value of the calf at birth This is given in equation (2)

$$\pi(\theta) = p(\iota, \theta)w(\iota, \theta)\, e^{-r\theta} - c\iota \int_0^\theta e^{-rt}\, dt \qquad (2)$$

The new variables are π = the present discounted profit of the fattening process, ι = a fixed bundle of daily inputs to the steer, independent of θ, c = the cost of the fixed bundle, ι, and p = the price per pound which may be obtained from a beef consumer for the steer at age θ Both the weight and the price of the animal are assumed to be functions of ι and θ, implying that the quality of the beef is reflected in the unit price received The inputs required consist primarily of feed, but conceptually may include all inputs such as labor, shelter, fences, machinery, and veterinary care[8] For the moment no attention is paid to the process of determining market prices, for these are given to the individual producer

---

[7] This might be called the "schmoo" model, except that the steer does not die voluntarily A steer is a castrated male valued only for its beef production

The first-order conditions for a maximization of π require that the producer select both the optimal slaughter age and the optimal input stream

$$\frac{\partial \pi}{\partial \theta} = e^{-r\theta}\left(p\frac{\partial w}{\partial \theta} + w\frac{\partial p}{\partial \theta}\right) - re^{-r\theta}pw - c\iota e^{-r\theta} = 0, \tag{3a}$$

$$\frac{\partial \pi}{\partial \iota} = e^{-r\theta}\left(p\frac{\partial w}{\partial \iota} + w\frac{\partial p}{\partial \iota}\right) - c\int_0^\theta e^{-rt}\,dt = 0, \tag{3b}$$

which yield

$$p\frac{\partial w}{\partial \theta} + w\frac{\partial p}{\partial \theta} = rpw + c\hat{\iota}, \tag{3a$'$}$$

$$p\frac{\partial w}{\partial \iota} + w\frac{\partial p}{\partial \iota} = c\int_0^\theta e^{-rt}\,dt = \frac{c}{r}(e^{r\theta} - 1) \tag{3b$'$}$$

At $\hat{\theta}$ the change in value due to changing weight and quality (unit price) is equal to the current interest foregone plus the cost of feeding Alternatively, dividing through by pw, the rate of weight gain plus the rate of price change due to aging is equal to the interest rate plus the cost per day of feeding the animal as a percentage of its total value Similarly, at $\hat{\iota}$ the present discounted value of the marginal net weight gain and price increase corresponding to a higher stream of inputs throughout the steer's life, less the present discounted cost of feeding the animal these inputs, must be zero

It is more important, however, to discover how the optimal $\hat{\theta}$ and $\hat{\iota}$ are affected by changes in the parameters faced by producers, that is, the

---

[8] The assumption that the input bundle is fixed is unrealistic The input bundle varies over the animal's life, and the animal's response to current inputs depends on the amount and timing of past inputs This becomes complicated mathematically, however, and has not been included in this analysis even though such effects are sometimes important Other factors can be more simply incorporated For example, marketing costs paid by the producer at the time of sale will tend to lengthen the slaughter age If these costs are fixed, and denoted by z, we have

$$\pi(\theta) = pwe^{-r\theta} - c\iota\int_0^\theta e^{-rt}\,dt - ze^{-r\theta}$$

and

$$\frac{\partial \theta}{\partial z} = -\frac{\partial^2\pi/(\partial\theta\,\partial z)}{\partial^2\pi/\partial\theta^2} > 0, \quad \text{for } \frac{\partial^2 \pi}{\partial\theta\,\partial z} = e^{-r\theta}(r) > 0$$

The producer reduces the present discounted value of the marketing costs by prolonging the time of sale During any period he slaughters fewer cattle, holding each to an older age Price expectations can be introduced by letting the price vector vary with time, $p(\iota, \theta; t)$, and the effect of climatic variation or disease on the animal's ability to convert feed into beef can be recognized by allowing shifts in the growth function, $w(\iota, \theta, \delta)$

price of beef, the cost of inputs, and the interest rate To determine this, the implicit function theorem may be used The function for probability is

$$\pi = f(t, \theta, r, c, p), \tag{4}$$

where the variables t, θ satisfy the subsidiary conditions

$$\frac{\partial \pi}{\partial \theta} = \Phi(t, \theta, r, c, p) = 0, \tag{5}$$

and

$$\frac{\partial \pi}{\partial t} = \Psi(t, \theta, r, c, p) = 0 \tag{6}$$

After writing $\hat{t}$, $\hat{\theta}$ as functions of r, c, p,

$$\hat{t} = \alpha(r, c, p), \tag{7}$$

$$\hat{\theta} = \beta(r, c, p), \tag{8}$$

α and β may be substituted for t and θ in Φ and Ψ Using the chain rule for differentiation, we may then solve for the unknowns,

$$\frac{\partial \hat{\theta}}{\partial p}, \frac{\partial \hat{\theta}}{\partial c}, \frac{\partial \hat{\theta}}{\partial r}, \qquad \frac{\partial \hat{t}}{\partial p}, \frac{\partial \hat{t}}{\partial c}, \frac{\partial \hat{t}}{\partial r}$$

Explicit numerical solutions for each of these unknowns can be obtained using data from farm management studies However, as only qualitative results were needed for this study, rough estimates of the price data and the growth functions pertaining to Argentine steers were used to obtain the following results These results will be valid as long as the rate of gain is declining in the vicinity of $\hat{\theta}$ and if the marginal return to increased inputs diminishes monotonically

$$\left.\begin{array}{l} \dfrac{\partial \hat{t}}{\partial p} > 0 \\[2ex] \dfrac{\partial \hat{\theta}}{\partial p} > 0 \end{array}\right\}$$

A rise in p increases the marginal value product of each input, increasing the optimal feed ration and the optimal slaughter age

$$\left.\begin{array}{l} \dfrac{\partial \hat{t}}{\partial c} < 0 \\[2ex] \dfrac{\partial \hat{\theta}}{\partial c} < 0 \end{array}\right\}$$

A rise in the cost of inputs reduces both the daily input and the optimal slaughter age Animals are fed not only less per day, but for a shorter period of time because they grow more slowly at any given age

$$\left.\begin{array}{l} \dfrac{\partial \hat{t}}{\partial r} < 0 \\[2ex] \dfrac{\partial \hat{\theta}}{\partial r} < 0 \end{array}\right\}$$

A rise in r reduces the daily feed inputs, because higher feed investment implies higher interest costs The increase in r also reduces the optimal slaughter age, as it increases the interest foregone at every age

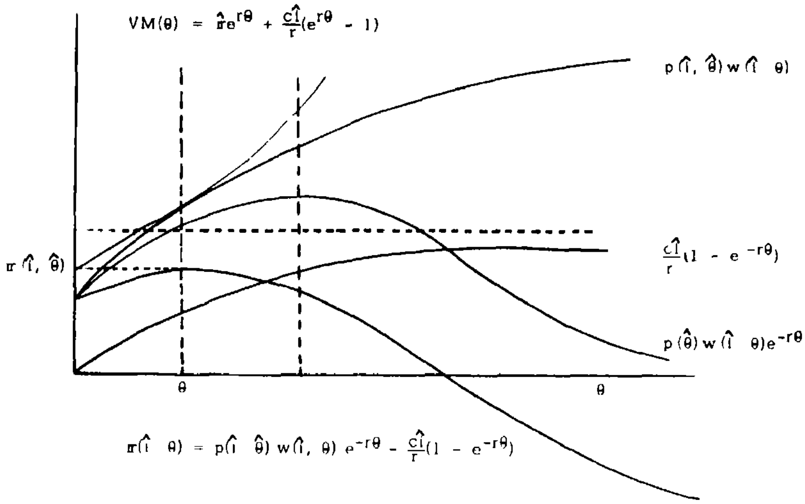


FIG 1

In particular, these results indicate that a negative slaughter response for steers is expected in the "short run." Fewer steers are slaughtered temporarily because a higher price causes them to be withheld. This, of course, is a *ceteris paribus* result.

Consider now the determination of the market prices of different-aged male animals, from calf to steer. We know that $\pi(\hat{\theta})$ represents the calf's value at birth, that is, $\hat{\pi}$ is the amount which if invested at interest rate r would have the same money value at time $\hat{\theta}$ as the finished steer, less the total feed costs compounded from their time of input to $\hat{\theta}$, at rate r. The calf has value because it is a "growing" machine.

$$\begin{aligned} \pi(\hat{\theta}) &= p(\hat{\imath}, \hat{\theta})w(\hat{\imath}, \hat{\theta})\, e^{-r\hat{\theta}} - c\hat{\imath}\int_0^{\hat{\theta}} e^{-rt}\,dt, \text{ and} \\ \hat{\pi}\, e^{r\hat{\theta}} &= p(\hat{\imath}, \hat{\theta})w(\hat{\imath}, \hat{\theta}) - \frac{c\hat{\imath}}{r}(e^{r\hat{\theta}} - 1) \end{aligned} \tag{9}$$

This result may be illustrated graphically while demonstrating another point as well (see fig. 1). In deriving the optimal slaughter age and input stream, it was assumed that producers faced known functions for the rate of gain and the rate of change of price per unit for each animal. The product of these functions would, if graphed as a function of age, yield the locus shown as $p(\hat{\imath}, \theta)w(\hat{\imath}, \theta)$. Given our previous simplified assumptions, slaughter occurs only at one age, $\hat{\theta}$, and because we assume perfect competition, the market value of the animal at $\hat{\theta}$ must equal the cost of producing the animal. This supply cost, reflecting the cost of feed inputs as well as the interest foregone on the value of the calf, can be easily obtained by rearranging equation (9) and is graphed as $VM(\theta)$ in fig 1. Slaughter occurs where $VM(\theta)$ is tangent to pw [9]

In fact, however, animals are slaughtered at many different ages This occurs both because some consumers are willing to pay a premium per unit weight for meat from either younger or older animals and because feed costs differ for different producers For the moment we ignore the latter factor and consider only the implications of the former Under our original assumption a calf's capital value dominates its slaughter value until age $\hat{\theta}$, and no calf will be slaughtered until this age Any producer wishing to sell a calf will find a buyer who will continue to feed the calf until age $\hat{\theta}$ However, any consumer wishing to purchase an animal at a different age could do so if he were willing to pay a premium price per pound Meat from animals slaughtered at other ages must bring a premium price because consumer prices must vary directly with $w(i, \theta)$ to ensure that the producer is fully compensated for the original value of the calf and the value of the embodied feed inputs, including interest To restate, if people are willing to pay $p(\hat{\theta})$ for meat from an animal aged $\hat{\theta}$, $p(\theta)$ must be greater than $p(\hat{\theta})$, $\theta \neq \hat{\theta}$ To show this, we return to the model where no feed inputs are required and consider the cost per pound of producing animals of different ages Take the case $\theta < \hat{\theta}$ Because $\dot{w}/w > r$ at this age, the animal must be worth more as a growing machine than as a consumption good If the calf is purchased for consumption, it must be at the price determined by the capital value This implies

$$p(\theta)w(\theta) = p(\hat{\theta})w(\hat{\theta})\, e^{-r(\theta-\theta)} \tag{10}$$

But now let $w(\hat{\theta}) = e^{g(\theta-\theta)}w(\theta)$, $g > r$ Then

$$\begin{aligned}\frac{p(\theta)}{p(\hat{\theta})} &= \frac{w(\hat{\theta})}{w(\theta)}\, e^{-r(\theta-\theta)} \\ &= \frac{w(\theta)}{w(\theta)}\, e^{g(\theta-\theta)}\, e^{-r(\theta-\theta)} \\ &= e^{(g-r)(\theta-\theta)},\end{aligned} \tag{11}$$

and $p(\theta) > p(\hat{\theta})$, $\theta < \hat{\theta}$ A similar proof can be used to show that $p(\theta) > p(\hat{\theta})$ for $\theta > \hat{\theta}$ Therefore, although the value of the animal itself, $VM(\theta)$, increases monotonically, the price per pound of the beef from the animal, $p(\theta)$, will have a U-shaped age profile, as shown in fig 2 The inclusion of feed costs into the model, so long as costs are the same for all producers, will increase the bowness of this profile

---

[9] As is clear in the mathematical formulation, the optimum slaughter age is a function of the relative beef/feed price, p/c, not the absolute price of beef

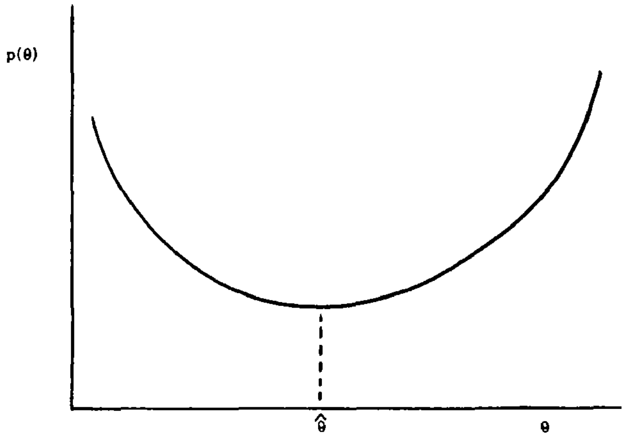


FIG 2

The derivation of the least-cost-per-pound age, and the premium paid for animals of all other ages, is useful for conceptualizing how consumer preferences can affect the age distribution of slaughter, both cyclically and secularly However, although consumers' preferences influence the age distribution of slaughtered animals by determining the premium consumers are willing to pay for beef from different-aged animals, these preferences have only an indirect effect on either the relative or absolute differences between the equilibrium market values of different-aged animals These relative prices will vary only slightly with changes in the price of beef and feed and the interest rate, the changes depending on the specific growth function of the animal A similar result holds for the absolute difference between the supply-determined market values of two steers of different age The supply-determined relative market values of different-aged steers, such as at age θ_1 and θ_2, are easily found, being given by equation (12)

$$RP_{\theta_1\theta_2} = \frac{VM(\theta_1)}{VM(\theta_2)} = \frac{\hat{\pi}\, e^{r\theta_2} + ci/r(e^{r\theta_2} - 1)}{\hat{\pi}\, e^{r\theta_1} + ci/r(e^{r\theta_1} - 1)}, \qquad \theta_2 > \theta_1 \qquad (12)$$

II

In the previous section simple economic models were employed to analyze the impact of various parameter changes on the market value, input level, and slaughter age of steers A similar analysis is now carried out for cows As may easily be understood, the principal characteristics distinguishing cows from steers, from an economic viewpoint, is the ability of the former to produce calves[10] Cows may produce beef either directly, by being fattened for slaughter, or indirectly, by bearing calves which may themselves be fattened for slaughter *This option is reflected by including an additional term in the profit equation analyzed previously* This term is the present value of the expected calf stream

$$\sum_{t=1}^{\theta} \frac{C(\iota, t)}{(1 + r)^t},$$

where $C(\iota, t)$ is the expected value of the calf born in year t, assuming the cow has been fed input stream ι throughout its life This expected value depends on the probability that the cow will have a calf in year t, the respective probability that the calf will be male or female, and the expected values of male and female calves in that year

Equation (13), the profit equation for females, includes three terms reflecting, respectively, the calf stream, the inputs required to maintain the animal, whose sum is equal to the value of a female animal at birth, and the present value of the beef available at the time of slaughter

$$\hat{\rho} = \sum_{t=1}^{\theta} \frac{C(\hat{\iota}, t)}{(1 + r)^t} - c\iota \int_0^{\theta} e^{-rt}\, dt + p(\hat{\iota}, \hat{\theta}) w(\hat{\iota}, \hat{\theta})\, e^{-r\theta} \tag{13}$$

This equation can be used to determine the optimal slaughter age and input stream for cows, as was done previously for steers It is important to note, however, that female calves have a distinctly bimodal optimal slaughter age This bimodality occurs because more female calves are born than are needed for replacement purposes for the breeding herd As a result, some female animals are slaughtered as fattened heifers at age θ_1, before they have begun to bear calves, and some are slaughtered only after their value as breeding animals has declined, at age θ_2[11] Female calves are essentially homogeneous at birth, and producers are therefore indifferent at the margin between retaining an animal for the breeding herd or fattening it for slaughter If the value of a female as a breeding animal rises relative to its value as a slaughter animal, some females formerly destined for slaughter will be withheld, and vice versa This switching will continue until an equilibrium is achieved[12]

---

[10] Biologically, male animals are as essential as cows to the breeding process Therefore, the profit equation for male animals ought to include a breeding term as well This term was deleted in the previous section to simplify the analysis The term does play an important role in some situations, however, and will be discussed in the next section

[11] Although cows may conceive until age 13, they are rarely retained in the breeding herd beyond age 9 because their teeth wear down, making it increasingly difficult to feed This lowers the probability of conception, which is sensitive to the cow's level of nutrition, and makes it increasingly difficult for the cow to suckle a calf Both factors decrease the expected value of the calf stream, prompting slaughter

[12] The models discussed in this paper focus on the partial-equilibrium behavior of producers facing exogenous changes in prices, although clearly such prices are endogenous to the economic system as a whole Without an endogenous solution, many relationships such as the bimodal slaughter distribution would not hold

As should be clear, an analysis similar to that carried out for steers can be used to show that the immediate response of both heifer slaughter and cow slaughter to an increase in the price of beef is negative The higher beef price, or lower feed cost, makes it profitable to feed to heavier weights and to retain animals for calf production

III

The models presented in the two previous sections can also be utilized to show that the magnitude of the slaughter response will differ for different types of animals

At any single point in time there is a fixed supply of animals in the herd for which there exist two types of demand consumer and producer As long as a producer is willing to outbid the consumer to retain the animal as a productive asset, the animal remains in the herd When the consumer wins, the animal is slaughtered Our interest is then to determine how the relative strength of the bidders' demand for different types of animals varies with exogenous shocks to the system, for example, devaluation and climatic variation

Some insight is provided by the partial-equilibrium models developed previously We examine first the *ceteris paribus* change in relative value between a steer ready for slaughter and a newly castrated calf, which follows an "exogenous" increase of 10 percent in the price of steer beef Assume that the slaughter age of the steer is unchanged in the new situation and that the value components of the profit equation originally have the following magnitudes, which are the approximate relationships prevailing for a steer calf in Argentina

$$pwe^{-r\hat{\theta}} = a = 5, \tag{14}$$

$$ci\int_{0}^{\theta} e^{-r\theta}\,dt = b = 4 \tag{15}$$

A steer calf then has value $\hat{\pi} = a - b = 5 - 4 = 1$, and we find that a 10 percent increase in the price of steer beef produces a theoretical rise of 50 percent in $\hat{\pi}$

$$\frac{\hat{\pi}^2}{\hat{\pi}^1} = \frac{1\ 1(5) - 4}{5 - 4} = 1\ 5 \tag{16}$$

The further away is the age of expected slaughter, the greater the change in the capital value of the animal

The effect is even greater in the case of female calves, or male calves which have not been castrated In either case, the option exists to retain the animal to an older age, but more important, an increase in the value of a calf increases the calf stream value of the female calf, or the stud value of the male calf, and the recursive effect of further calf price increases would be carried on indefinitely if there were no damping force Thus, the proportional "instantaneous" increase in the value of a "breeding" calf should be substantially greater than either the original rise in the price of steer beef or the rise in the value of a castrated male calf

The "instantaneous" increase in value which is reflected in these models is, of course, a partial-equilibrium result These models take no account of the fact that other adjustments, perhaps quite rapid ones, will occur in response to a beef price increase For example, the change in the value of the steer calf discussed above reflects the change in the value of the animal as a capital good As this value rises, producers will respond by retaining more such animals to be used for future production, and will reduce the number currently slaughtered The resulting reduction of current slaughter will increase even more the current price of beef but will also increase the future supply of beef, thereby lowering the expected future price of beef and perhaps increasing the cost of feed As the capital value of an animal depends on expected as opposed to current prices, the movement of expected prices will dampen, at least at some point, the tendency for the relative prices of animals to change [13]

The process will be facilitated by the fact that beef from different animals is highly substitutable in consumption Thus, as the relative prices of certain types of animals begin to increase, the consumption of these animals is reduced Market prices are constrained by the high price elasticity of consumer demand across animal categories, and this constraint allows producers to bid away more easily those animals with more sensitive capital prices [14]

This consumption constraint is important in two other respects as well First, because it limits the relative price variation, the differential effect of a price increase indicated in the partial-equilibrium models will tend to be reflected in the slaughter response of the different animal categories The slaughter response also depends on the relative availability of the animals, but the consumption constraints play an important role Second, the limitation on the relative price variation allows the use of a single price, such as the price of 2-year-old steers, as the price variable for all types of cattle in a disaggregated econometric model, without great loss of accuracy. This is helpful because it is difficult to identify the consumer demand for each type of animal

---

[13] For simplicity, I have assumed in the model to this point that the price of beef as a function of age and of food inputs is given The same results hold if producers' expected prices are determined similarly It is necessary, however, to consider the formation of price expectations if the dynamics of the adjustment process are to be clear The capital values of certain animals may be decreasing even at a moment when their current slaughter value is increasing This movement, which prevents any price movement from being cumulatively destabilizing, requires that price expectations take into account future supply and demand, rather than naively extrapolating current price movements

[14] Only when no more cows and heifers can be withdrawn from slaughter will the "tie" via consumer demand between their prices and those of other animals be broken Under normal conditions, given the number of cows and heifers sent to market each year and the cost of making large sudden changes in the size of the breeding herd, this does not occur

Although each animal category should exhibit a negative short-run response to a price increase, the information contained in the immediately preceding paragraphs suggests an expected ranking of slaughter elasticities of the different categories The differences in these slaughter elasticities will reflect both the sensitivity of the value terms in the profit equation, as shown in the partial-equilibrium models, and also the relative availability of each animal category The degree of instantaneous impact of a parameter change on the capital value of the animal depends on the expected time lapse before slaughter and on the presence of the breeding term with its recursive effect The relative supply is also important, however, for the elasticity of slaughter response refers to the percentage change in the number slaughtered The larger is the number of animals of a given category relative to the number needed to satisfy the increased herd demands, the lower the expected elasticity This point, which may appear tautological at first, may be illustrated by considering again the role of the breeding term in the profit equation for male animals

In principle, there is no difference in the equation for male and female calves at birth A male calf also has a bimodal optimal slaughter age, for it too can be fattened to be slaughtered for beef or retained to enter the herd as a breeding animal As each male calf theoretically has the potential to do either, each male profit equation should contain a breeding value term Accordingly, the value of a male calf at birth may be as sensitive as that of a female calf, the value of a male calf becomes less so only after castration Following castration male calves' values will be less sensitive because their productive value then depends only on their ability to convert feed into beef, but this difference is an artificial factor introduced by producers [15]

The price elasticity of female slaughter is, however, normally greater than that of males because there are fewer female animals available relative to the need for the breeding herd One bull may service many cows, and hence fewer bulls than cows are required for the breeding herd Because male and female calves are born in roughly equal numbers, the proportion of male calves which must be retained from slaughter each year for replacement purposes is considerably smaller than the proportion retained of female calves Similarly, when the size of the breeding herd is changed, the proportion of male animals switched from prospective slaughter to the breeding herd is generally much less than the proportion of female animals because there are many more male animals destined for slaughter and many fewer required for any desired change in the breeding herd. The availability factor immediately produces a difference in the slaughter elasticities of the two types of animals

---

[15] The decision to fatten or to retain for breeding is generally taken quite early in the case of males because very few males are needed for breeding, and also because the costs of castration are lower, the earlier the age Castration sacrifices the value of the breeding component and generally inhibits growth, but it does make the animal more docile, thus reducing management costs and the likelihood of future injury

It is therefore difficult to generalize the expected elasticity of slaughter response One must consider the difference between the animal's actual and expected slaughter ages, its breeding potential, and the normal distribution of slaughter A convenient rule of thumb suggests that female animals should have a higher slaughter elasticity than males, and younger animals, a higher elasticity than older For example, while male calves, even prior to castration, should demonstrate a lower elasticity than females (because of the lower relative demand for the breeding herd, not the absolute lack of breeding potential), and while both male and female calves should have a more elastic slaughter response than either steers or cows, bulls should demonstrate a more elastic response than steers, despite the fact that the bull is generally an older animal Further, it is not strictly necessary that heifers exhibit a higher slaughter elasticity than yearling steers, for although some heifers may be switched from slaughter to the breeding herd, those which are not cannot be withheld for further fattening very long It is not profitable In contrast, yearling steers, even though lacking breeding potential, are frequently withheld for fattening for a longer period of time

IV

The previous models suggest that the immediate slaughter response is negative for all categories This does not necessarily imply that the estimated beef/feed price coefficients in the slaughter equations of an econometric model will be negative for all categories First, an increase in the price of beef which is not expected to last could lead to increased rather than decreased slaughter in the very short run It is necessary to differentiate between the response to an expected price and the estimated coefficient on a past or existing price variable This is a well-known problem

Second, slaughter response is quite different from production response The attempt by producers to increase production requires a reduction in slaughter in the short run, and the stronger is this attempt, the sharper is the drop in slaughter Production, however, will increase eventually to allow greater slaughter As the period of observation grows longer, the (net) slaughter response becomes less negative, eventually becoming positive The estimated sign of the beef/feed price coefficient in the slaughter equation therefore depends entirely on the length of the period of observation, that is, a quarterly model is more likely to exhibit a negative beef-price slaughter elasticity than an annual model[16] The rapidity with which the buildup in stocks is reflected in a higher slaughter flow is likely to vary across categories

Third, the fact that animals can pass through several categories during 1 year means that "switching" caused by price changes can affect the estimated price coefficients According to Argentine definition, an animal is a calf from birth to 9 months, a yearling from 9 to 18 months, and a steer from then until slaughter A calf aged 8 months at the beginning of the year theoretically could be slaughtered during the year in any of the three categories Thus a price increase which causes all animals to be fed to heavier weights may cause some animals to be withheld just long enough to be slaughtered in a category different from that in which they would have been slaughtered before the price increase This effect could make it incorrectly appear that the animals in older categories had a positive price-slaughter response[17]

Finally, because a price increase can alter the age distribution of the slaughtered animals within each category, a price increase can also change the amount and type of beef produced even if the number of animals slaughtered is not affected This probability suggests the need to include in the model equations estimating the average slaughter-weight of the different categories if a good prediction of total beef production is desired

V

The micro-models developed previously can also help explain the distribution of production activities among regions or among countries once the assumption of equal input costs for all producers is dropped For example, assume that the cost of transportation for dressed beef is negligible but is considerable for live animals, and that consumers are willing to pay only very small premiums for beef from different-aged animals This makes the consumption value of the animal essentially a function of weight Thus, in regions where feeding costs are relatively higher, the capital value of calves will tend to be lower, at least up to some age θ_i Because

$$\hat{\pi} = pwe^{-r\theta} - ci \int_0^\theta e^{-rt}\, dt,$$

for any given p and r, a higher c will be associated with a lower $\hat{\pi}$ Further, as $VM(\theta) = \hat{\pi} e^{r\theta} + ci/r(e^{r\theta} - 1)$, a higher c results in a lower $VM(\theta)$ during the early part of the animal's life, $\theta < \theta_i$, and a higher $VM(\theta)$ after some point in time, $\theta > \theta_i$ This may be graphed as in figure 3

---

[16] Recent work by Nores (1972) confirms this

[17] Another reason for encountering a positive price coefficient has been suggested by Yver (1971) in subsequent work and is discussed later

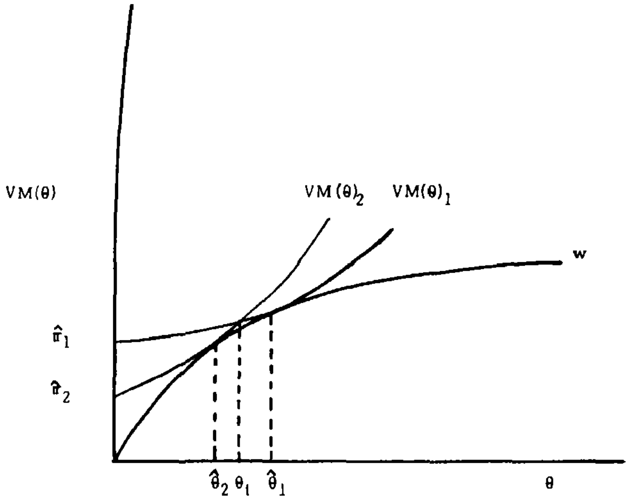


FIG 3

Considering regions like Europe and Argentina, we can see why veal in Europe is absolutely cheaper per pound than in Argentina, although mature beef is much cheaper in Argentina. If transportation costs were zero, calves born in Europe would have value $\hat{\pi}_1$ at birth and be shipped to Argentina to be fattened to age $\hat{\theta}_1$. As transportation costs are not zero, the value of calves is $\hat{\pi}_2$ and they are fattened until slaughter in Europe. But they will not be fattened past θ_1, because after this age imported beef of the same quality is cheaper. The cost per animal of varying ages in Europe therefore should have the shape of the envelope in figure 4. As can be seen, the existence of many producers located in regions with

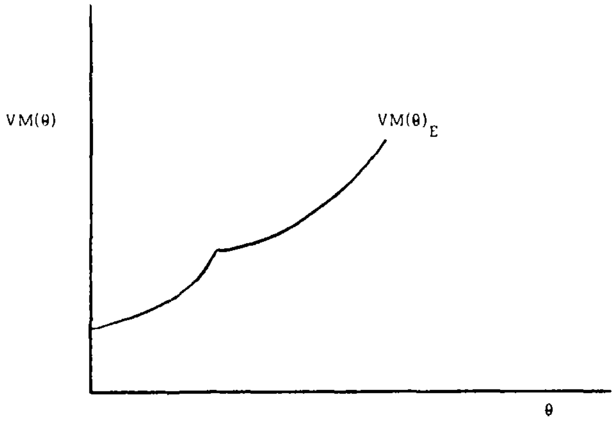


FIG 4

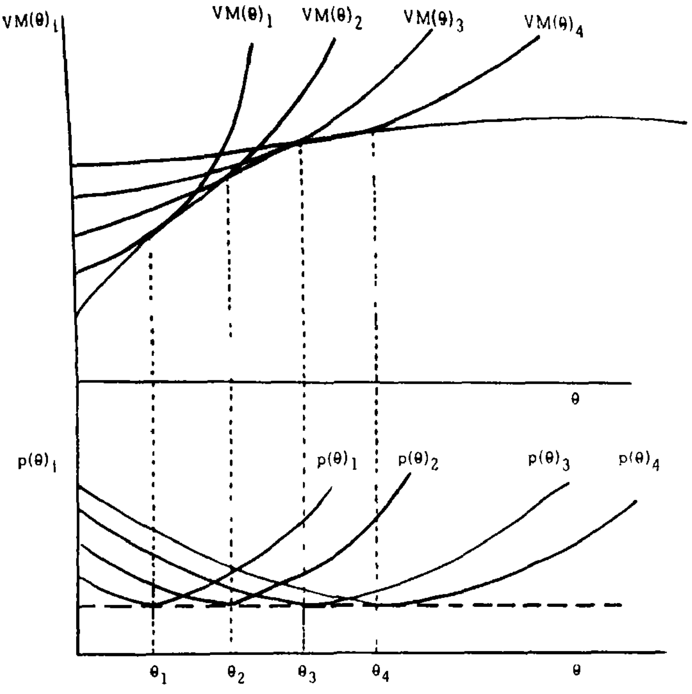


FIG 5

different feed costs implies that the observed relative cost per pound of beef aged θ_i will vary less than if there were only one producer Further, the absolute cost of beef from younger animals is cheapest in the higher-feed-cost regions [18] This relationship is shown in figure 5, where it is assumed that there is strict consumer indifference among beef from animals aged θ_1, θ_2, θ_3, and θ_4 Each country is shown to be relatively more efficient in the production of beef from a certain range of animal ages

The regional location of production activities within a country is determined similarly Producers with different feed costs will choose different parts of the production process For example, breeding operations will usually take place in areas where the opportunity cost of feed is cheap, that is, where the cost of maintaining the cow year round is less than the value of the calf at birth Because all calves will have the same value at birth in a unified market, it will not be profitable to maintain breeding herds in high-cost feed areas unless producers there are more efficient, that is, unless their herds have higher calving rates and lower mortality rates than herds elsewhere [19]

---

[18] Thus, part of the observed European "preference" for veal results from veal's lower relative price therein

Breeding may also take place as a subsidiary or complementary activity, in some areas cows are maintained primarily for milk production In this case the profit equation should include a component reflecting the present discounted value of the future milk stream

$$P_m \int_\theta^\theta m(i, t)\, e^{-rt}\, dt,$$

where $m(i, t)$ is the quantity of milk produced by a cow aged t, fed inputs i, and P_m is the price at which this milk can be sold In these areas the milk component must compensate for the lower net value of the calf stream, which is implied by the (usually) much higher feed costs [20]

The fattening process may enjoy similar regional specialization, depending on the opportunity cost of feed in different regions and the length of the period for which it is available Upon weaning, calves are usually sent to fattening regions where feed suitable for rapid weight gain is available After this the animals may be sent to market or sent to lower-cost grazing lands for final finishing Whether any individual animal is sold to slaughter or to further fattening depends on the current market price for slaughtered beef, whether any producer has feed available at a cost low enough to permit further profitable feeding of this animal, and transaction and transport costs If current prices for feeder animals and expected future prices for finished animals are in the correct ratio, producers who have low-cost feed will decide to purchase animals to utilize it and then sell them at the end of the period of feed availability, either to other producers for further fattening or directly to slaughter In many areas there will be feed during a particular part of the year which can be fed to cattle quite cheaply, for example, winter wheat can be grazed for several months without damage to the crop, and the wheat stubble may be grazed following the harvest Hence wheat and cattle may be complementary during these periods, even though they may be competitive activities in other situations

---

[19] In Argentina the major breeding area is the Salada River basin, where drainage and land quality are relatively poor Sufficient pasture for the cows and calves is usually available during the crucial periods of the year, but grain and/or forage crops for year-round fattening normally cannot be grown here Breeding herds are maintained elsewhere, however, and there is evidence that these producers achieve higher calving rates See Jarvis (1969, chap 8)

[20] There is no reason that the capital values of male and female calves at birth should be identical For dairy animals male and female calf values will differ substantially, and the behavior of producers toward calves of different sex should also differ sharply Male calves are produced solely as a by-product of the attempt to obtain female calves for future milk production, and male calves will be slaughtered shortly after birth

VI

Steers are now slaughtered at younger ages in Argentina than was true some years ago There are several reasons for this First, the effective interest rate (rate of discount) faced by farmers has increased, for farmers have become more sensitive to alternative investment opportunities This should reduce $\hat{\theta}$, the least-cost-per-pound age, although the effect would likely be small Second, the relative cost of feed inputs has also risen, that is, the cost of pasture and forage has risen relative to the price of beef, especially as the grain yields for land have risen Because younger animals convert feed into beef more efficiently than older animals, a rise in the relative price of feed will tend to reduce the slaughter age Thus both cost factors considered by this model have tended to reduce the slaughter age of animals in Argentina

The slaughter age also depends on the extent of the premium which consumers will pay for beef of various ages Here too the trend in recent years seems to have favored the slaughter of younger animals Many consumers now actually prefer the leaner beef from a younger animal and are quite unwilling to pay the premium which was once received for older, fatter beef Lucio Reca has suggested that this is due not only to a change in tastes for the beef itself but also to the growth of the vegetable oil industry, which has provided an attractive and preferred substitute for animal oils As one of the primary sources of the latter used to be fatty beef, the demand for fat as a complementary product of beef has decreased [21]

New breeds of cattle and better pasture management have also permitted higher growth-conversion rates at younger ages, and an earlier "leveling-off," leading to slaughter at an earlier age

VII

The previously developed micro-theory was used to specify the equations of an econometric model of the cattle sector The model seeks to explain the changes in the herd size, the number and weight of the animals slaughtered, domestic consumption, exports, and the price of beef in reaction to exogenous shocks such as climatic variation, devaluation, a change in the level of domestic income or population, and so forth Because of the scarcity of space only a cursory description of the model and the empirical results are presented, a more detailed discussion is contained in Jarvis (1969, chap 5–9)

---

[21] Personal conversation Reca was then director of the Department of Agricultural Economics, Ministry of Agriculture, in Argentina

The model contains 21 equations, including six identities There is for each of the six animal categories an equation to estimate the number slaughtered and an equation to estimate the average weight of those slaughtered The number of calves born is also estimated, but the number of animals in each of the other five categories in year t is given identically by making use of the constraint relating the herd stocks in 2 different years with births, slaughter, and natural deaths Two equations estimate domestic and foreign consumption (export) demand, each in terms of tons of beef, and a final identity equates beef produced from slaughter with total consumption[22]

The model satisfies the posterior rank-and-order conditions for identification, all the equations are overidentified The model was estimated using ordinary least squares and instrumental variables The results were highly similar, and only the latter are presented here Moreover, as this paper is concerned primarily with the behavior of slaughter as a test of producer behavior, only the slaughter equations are given An aggregate slaughter equation is provided for comparison See table 1

The slaughter equation has essentially the same form for each category

$$S_t = \alpha H_t + f(x^1, \quad , x^n) + E_t, \tag{17}$$

where the function f contains both current and lagged variables This approach assumes that there is a permanent or normal component and a transitory component to observed slaughter.[23] In equilibrium a constant proportion of the herd, or category, is slaughtered each year This number, however, may be increased or decreased depending on the desires of producers, which in turn depend on the level of certain parameters These parameters, such as the current price level or current climatic conditions, affect producer expectations and thereby the size of the desired future herd

Because the coefficient on the stock variable should approximate the average number of animals slaughtered each year, the other independent variables are included as deviations from their mean Thus, in a "normal" year, the "normal" proportion is slaughtered However, the proportion of the stock slaughtered in certain categories has varied secularly, and it was necessary to allow for this occurrence The secular increase in the calving rate has provided a rising number of heifers relative to the replacement needs of the breeding herd, and a growing proportion of the heifer stock has been slaughtered over time Similarly, a growing preference for leaner beef has increased the proportion of yearlings slaughtered And finally, it appears that a greater, although still small, proportion of bulls is being raised specifically for beef production, thus increasing the rate of slaughter of the bull herd.[24] These effects were included by making the "normal" proportion slaughtered a function of time

---

[22] It was assumed that the margin between producer and retail prices is constant so that only one price is endogenously determined in the system

[23] A discussion of the interpretation of this form versus that of a lagged stock adjustment model is contained in Jarvis (1969, chap 6) Several nonlinear formulations of this model were estimated, but each was less satisfactory than the form shown here

$$\alpha H_t = (\alpha_0' + \alpha_1' t)H_t = \alpha_0' H_t + \alpha_1'(tH_t) \qquad (18)$$

The coefficient α_1' was significant for several of the categories, and after this adjustment α was in fact equal to the mean number slaughtered for each category

Consider also the specification of the price coefficients It is assumed that producers respond to an expected price when making their slaughter decisions and that this expected price is a function of current and past prices and the current rate of change of price The higher the price in any year, or the greater the current rate of change of price, the greater is the expected price and the greater the desired future herd Several alternative specifications for the price variables were tried, as it was not certain how producers might form their expectations, or whether the same variables would have the same effect in different equations The relevant expected price for each category is the price expected to prevail at the time the animal, or its product, will be sold The producer need be concerned in the case of steers only with the price which will hold in the immediate future, whereas the relevant price for a breeding heifer is the average price which will prevail over the period in which her calves will be born Thus, the forms of the lagged distribution need not be the same for all categories Nonetheless, except in two cases they are highly similar

The three alternative specifications estimated are based on the respective assumptions that the relevant expected price is a function of (1) current and past prices with no weight being given to the current rate of change, (2) the current price, past prices, and the current rate of change, (3) past prices and the current rate of change The results were as follows For every category the rate of change of price was significant in formulation 2, which produced uniformly better results than formulation 1 Formulation 3 did, however, provide better results than formulation 2 in three cases, steers, yearlings, and heifers That is, the rate of change of price was always a significant determinant of slaughter, and for three categories, those male animals which are destined to be slaughtered in a reasonably short period of time, better results were obtained when only the current rate of change of price was included and the current price level excluded Following the estimation of these three general formulations several additional equations were estimated for each category to

---

[24] Producers may also be culling their bulls more systematically and slaughtering breeding bulls at lower average ages

TABLE 1

SLAUGHTER EQUATIONS, 1943/44–1965/66

	H_t	tH_t	P_t/P_{t-1}	P_t	P_{t-1}	P_{t-2}	P_{t-3}	P_{t-4}	P_{t-5}	ARL_t
S_t	0 1650 (18 35)	0 0016 (3 50)	−30,735 (5 94)		−19,520 (8 42)	−4,771 (3 15)	3,399 (2 14)	4,989 (4 14)		296 (3 08)
NS_t	1 0745 (17 70)	−0 0055 (1 63)*	3,623 (1 41)*		4,620 (0 43)*	3,300 (3 46)	4,168 (4 54)	3,068 (4 70)		18 (0 40)*
VS_t	0 0099 (17 2)	0 0008 (2 45)	−3,453 (1 40)*	−5,854 (4 54)	−2,049 (2 49)	−98 (0 09)*				146 (4 01)
YS_t	0 2210 (45 6)		−8,243 (6 16)		−2,424 (4 07)	−1,886 (4 79)	−1,375 (3 69)	−890 (2 51)	−432 (1 80)	96 (3 96)
VQS_t	0 1050 (2 09)	0 0072 (2 91)	−11 026 (4 90)		−4,696 (4 82)	−2,023 (3 12)	−217 (0 35)*	722 (1 21)*	794 (1 97)	−24 (0 57)*
TS_t	0 0672 (24 8)		−4,537 (2 02)	−2,509 (2 17)	−1,307 (1 76)	−471 (0 47)*				167 (5 70)
BS_t		0 0062 (50 0)	367 (2 26)	−292 (4 56)	1 790 (0 04)*	178 (3 29)	236 (4 33)	177 (4 74)		−1 (0 43)*

Table 1 (*Continued*)

	EB_{t-2}	C_t/C_{t-1}	C_t	C_{t-1}	C_{t-2}	C_{t-3}	C_{t-4}	C_{t-5}	$\bar{R}^2$	d	SER (δ)
S_t		−15,337 (2.51)		−340 (4.06)	−9 (0.15)*	158 (3.13)	161 (4.50)		0.92	2.44	3,844 (.045)
NS_t	−33,510 (3.19)		−97 (2.97)	−26 (1.23)*	14 (0.61)*	23 (1.25)*			0.85	2.25	2,064 (.051)
VS_t			−36 (1.53)*	−37 (1.94)	−25 (1.38)*				0.86	1.69	1,505 (.062)
YS_t		−2,995 (1.86)		−1 (0.05)*	−8 (0.57)*				0.76	1.18	1,037 (.097)
VQS_t		−6,299 (2.19)		−7 (1.92)	12 (0.39)*	64 (2.48)	79 (3.72)	58 (4.31)	0.86	1.68	1,692 (.133)
TS_t			−50 (2.42)	−36 (2.61)	−20 (1.41)*				0.62	1.73	1,364 (.183)
BS_t				−4 (2.10)	4 (4.11)	8 (6.51)	6 (6.50)		0.95	1.94	137.0 (.083)

NOTE.—The dependent variable in eq. (1) is S_t, the total number of animals slaughtered in year t. The dependent variables in eqs. (2)–(7) refer to the slaughter of individual categories: steers, NS_t, cows, VS_t, yearlings, YS_t, heifers VQS_t, calves, TS_t, bulls, BS_t. Correspondingly, among the independent variables H_t refers to the number of animals in the herd in year t of the respective category considered by each equation, and similarly for tH_t, where the respective herd stock is multiplied by a time trend. P_t/P_t = percentage change in price of beef relative to an index of grain prices. ΔRL_t = absolute change in rural labor force in year t. EB_{t-2} = percentage of beef slaughter exported to Great Britain in year $t-2$. C_t/C_t = percentage change in climatic index in year t. (Different climatic variables were used for each of the different animal categories. Weather variations in each of the different locations were weighted by the proportion of the relevant category maintained in the region of each weather observatory.) All independent variables have been taken as deviations from their respective means, except the herd stock variables, in order for the herd stock coefficients to reflect the mean proportion of animals slaughtered each year. The price and herd variables were treated as endogenous. Instrumental variables were used to estimate the equations: the instruments were: climatic conditions in current and past years, level of domestic population, domestic per capita income, change in rural labor force, amount of devaluation in years t and $t-1$, an index of grain prices in $t-1$, wholesale price index for nonrural goods in year $t-1$, current level of beef export taxes, percentage of herd vaccinated against hoof and mouth disease, percentage of beef exports destined for Great Britain, size of herd in years $t-1$ and $t-2$, and dummies for the years of World War II. The t-statistics asymptotically correct are in parentheses below the coefficient estimates. For each equation df = at least 15. All coefficients are asymptotically significantly different from zero at the 5 percent level using a one-tailed test, except those marked with an asterisk. $\bar{R}^2$ = multiple correlation coefficient adjusted for degrees of freedom. d = Durbin Watson statistic. SER = standard error of the regression, δ = coefficient of variation of the estimated equation in parentheses below. Price and climatic coefficients are derived from a second-order polynomial lagged distribution. The program used was that developed by Hall (1967). In each equation the lag distribution on prices and climate was constrained to equal zero in year $t-i$, which follows the year of the last price coefficient shown in the table.

determine the precise specification of the price and climatic lag distributions which provided the best results A discussion of these results and their sensitivity (low) to changes in specification is included in Jarvis (1969), only the preferred results for each equation are presented in table 1

The econometric results are quite satisfactory Each equation explains well the variation in a slaughter of the respective category, and, more important, the coefficients of the independent variables are highly significant and consistent both in sign and magnitude with the previously developed theory A negative coefficient is expected on each of the price variables in the slaughter equations As can be seen, this is true for every category except steers and bulls, in these equations the coefficient on the price in year t is negative, but the other price variables have positive coefficients Several possible explanations for this are suggested subsequently The equations in table 1 indicate that the annual rate of slaughter of animals of every category but one, steers, is reduced temporarily by an increase in price For five categories the net impact of the current rate of change of price and price level on slaughter is negative, and greatest, relative to the mean number of animals slaughtered, for calves, heifers, yearlings, cows, and bulls in descending order These results are consistent with the theory developed previously The impact on steer slaughter is probably positive but extremely small

One might expect that the effect of the current rate of change of price would be strongest for those animals destined for slaughter in the near future, that is, the extrapolation of the current rate of change might hold for short periods only, with longer price expectations being based on an average of past prices Empirically there is some support for this The relative impact of the current rate of change of price is, in descending order, strongest for heifers, yearlings, calves, cows, bulls, and steers, although the latter two are positive The animals most strongly affected are then heifers and yearlings, animals which are destined for slaughter within a shorter period of time than are calves or cows The normal proportion slaughtered of the heifer and yearling steer stocks is small, and it appears that producers are easily able to retain even greater numbers Yearlings are retained for further fattening and are held to steer age in many cases, while heifers are retained for both additional fattening and breeding

As noted above, the sign pattern of the price coefficients in the steer and bull equations appears at first to be contrary to expectation The coefficient on the rate of change of price is positive in both equations, as are the coefficients on the lagged price distribution after year $t - 1$ The latter pattern appears unreasonable because it implies that the expected price is an inverse function of past prices However, the sign pattern could be explained by either of two effects First, a price increase inducing producers to withold yearnings from slaughter in greater numbers may cause a temporary change in the age distribution of the steer herd This change can affect the percentage of the steer category slaughtered in future years as the adjustment works itself out In the model used, the coefficient on the stock variable is not allowed to vary cyclically, and the effect of changing the proportion of the stock slaughtered over the cycle is forced onto the lagged price variables Second, an enduring increase in the beef/grain relative price ought to increase the number of steers slaughtered relative to the number of yearlings, and correspondingly lower and raise the respective proportion of the two categories slaughtered annually If so, the positive coefficients on the lagged price variables in the steer slaughter equation could reflect this effect Such an interpretation is supported by the long and significant distributed lag on the price coefficients, of opposite (and expected) sign, in the yearling and steer slaughter equations

The positive sign on the rate of change of price in the steer equation has several possible explanations First, if the price effect causes yearlings to be withheld, as it does, but only for a moderate time period, the withheld yearlings may become steers within the year and thus be slaughteren in a different category This could make it appear incorrectly that the steers existing at the time of the price increase were slaughtered rather than withheld The question, of course, is whether enough such withheld yearlings could be slaughtered as steers within the year to explain the observed coefficients [25]

Another explanation has been offered by Raul Yver (1971), who suggests instead that the theory developed in this paper can be made entirely consistent with an actual increase in the slaughter of steers, that is, the econometric result need not stem from measurement difficulties He argues that if producers face a short-run feed constraint, they will be unable to increase the herd in the short run as much as they would like to do in the long run Their desire to retain animals of all ages wil cause a rise in the opportunity cost of feed, which in turn will prompt the slaughter of some The animals most likely to be affected will be those which are nearing their time of slaughter, such as steers, for the capita values of animals with longer productive lives will be less sensitive to a short-run change in the cost of food

---

[25] The number of yearlings being withheld and slaughtered as steers in the same year need not be large, for the absolute magnitude of the coefficient on the rate of change of price in the steer equation is only one-half that of the corresponding coefficient in the yearling equation Moreover, the coefficients on the rate of change of price in the equations explaining the average slaughter weights of steers and yearlings (not presented here) are positive, but insignificant The slaughter-weight results suggest that there is no significant change in the average age distribution of either category, as would be the case if the early slaughter of steers were prompted by a price increase Yver (1971) cites data on the average slaughter weights of the different categories, including their cyclical variation, to suggest that withheld yearlings could not reach steer weight within a period of 1 year The data given, however, provide evidence only for the mean, not the whole distribution of slaughter weights, and hence do not prove the point Moreover, the yearlings withheld are likely to be fed more intensively than the average yearling, thus gaining more rapidly The yearlings will be retained by producers newly entering the fattening activity, using land which was formerly dedicated to grain production and which is of high quality Their better feed suggests that these animals will grow more rapidly, and should be slaughtered younger than were steers previously, especially as they will also face a higher opportunity cost for feed

Yver's explanation is ingenious and plausible, and I am sorry not to have thought of this possibility earlier I recognized that feed costs would rise as the demand increased, but had not considered that this could be sufficient to actually increase the slaughter of certain categories in the short run As discussed shortly, this may be the situation for bulls as well as steers However, the feed constraint could not last very long, as additional land could be made available for pasture and forage if larger herds were desired Therefore the higher rate of slaughter as reflected in the lagged price coefficients could not be due to a feed constraint but must be caused by something else More important, Gustavo Nores (1972) has estimated a quarterly model of the Argentine cattle sector in which the coefficient on current quarter's price in the steer equation is negative and highly significant Whether yearlings cross categories, or whether an increase in the opportunity cost of feed occurs with some lag, inducing producers to sell more steers in the intermediate run, the immediate slaughter response of steers appears to be negative[26]

The price coefficients in the bull equation are remarkably similar to those in the steer equation but somewhat more puzzling, as bulls being breeding animals, were expected to behave similarly to cows A suggested explanation is that a large number of slaughtered bulls are not cull bulls but bulls raised specifically for slaughter Castration reduces the rate of growth of an animal, and some producers are willing to forego the advantages of easier herd management to obtain greater efficiency in feed conversion This willingness has increased over time, also explaining why the bull stock coefficient in the slaughter equation has risen secularly[27] Thus, the animals in the bull category are more heterogeneous than are the animals composing other categories, and the price coefficients in the bull equation reflect the net effect occurring from the withholding of bulls for the breeding herd and the increased slaughter of those bulls fattened for that purpose The apparent increased slaughter in the latter supports Yver's argument because the average slaughter weight of bulls declines significantly in response to a price increase, suggesting the slaughter of younger and lighter animals in greater proportion at this time It is difficult to argue that this effect is created by animals changing categories

---

[26] When I estimated the steer slaughter equation without including the current rate of change of price, the coefficient on the current price was invariably negative, though small in magnitude The use of a polynomial distributed lag could have forced the negative coefficient on the price in the current period in my results, but Nores does not use a constrained distributed lag

[27] James H Cothern, extension economist, University of California, Davis, suggests that this has occurred in the United States, and I assume it may be true in Argentina as well

An aggregate slaughter equation was also estimated for comparison with the individual-category slaughter equations It is presented in table 1 This equation performs very satisfactorily, but it does not yield the detailed insights into the behavior of producers, and the information covering the changing composition of slaughter, which the individual-category equations provide, nor is the predictive ability of the aggregate slaughter equation, in the limited context in which this is understood here, as great By means of comparison, the squared correlation coefficient between the sum of the predicted values for the individual slaughter equations and the actual total slaughter is 976, and the regression coefficient of the actual total slaughter on the sum of the predicted slaughter from the disaggregated equations is 996 Each of these results suggests that the set of disaggregated equations explains the variance in slaughter better than does a single aggregate equation [28] The same result holds when the average-slaughter-weight equations are considered Disaggregation provides more information about the operation of the cattle sector with respect to producer behavior and also provides a more accurate indication of the quantitative and qualitative changes in the supply of beef

The effects of several other variables are briefly noted First, the lag distribution on the climatic variables is similar in sign and form to the distribution on price variables Climatic improvement reduces the slaughter of every category except bulls in the first year, with the strongest impact in the heifer and yearling equations Good climate also reduces the slaughter of steers and calves, and if conditions improve for more than 1 year in a row, cows are retained as well

The variable measuring the percentage of total exports to Great Britain reflects the fact that the British traditionally demanded fatter and therefore older steers The production of these animals increased the average age of slaughter, thereby decreasing the proportion of the steer stock reaching slaughter age each year

Producers' reactions toward risk can be measured by including in the slaughter regression equations the variability of those product and input prices which affect the relative profitability of rival activities (see Behrman 1967) In this study the variability of cattle and grain prices in real terms had no significant effect on the slaughter of cattle, and, for simplicity, they have been omitted here Government intervention in the tenancy and labor markets was important, however This intervention made grain production more risky than cattle production by increasing the threat of expropriation and of labor strikes during the harvest There was an eventual breakdown in the traditional tenancy system, accompanied by the exodus, voluntary and forced, of many rural workers Producers discarded their previous rotational system and switched out of field crops into cattle production, which was less labor intensive The slaughter of certain types of animals was reduced during this period as new producers entered production The variable for the change in the rural labor force is a proxy for tenancy disruptions in the agricultural sector[29]

---

[28] The root mean square error of the series predicted using the disaggregated equations is 2,356, while the mean number of animals actually slaughtered is 98,500 Thus, the mean error is less than 2 4 percent, and the residual of largest magnitude, 4,430 in 1964/65, involves an error of only 4 8 percent

Although the theoretical model seems to imply an important role for the interest rate, the latter actually plays a relatively small role Cattle raised for slaughter have a short life span and their slaughter age is much more heavily influenced by changes in the price of beef and food inputs (including climatic effects) than by changes in the rate of interest Moreover, changes in the real rate of interest charged on agricultural loans have been dominated by changes in the rate of inflation, which averaged 24 percent during the post-World War II period When the rate of inflation was included in the slaughter equations, its coefficient was insignificant The residual pattern indicates, however, that large discrete changes in the rate of inflation may have had a brief impact on two occasions[30]

VIII

Among the implications of the study, the following stand out First, the micro-models in this paper indicate that an econometric model should be disaggregated to obtain a clear understanding of producer behavior and improved estimates of future beef output Animals of different age, sex, and breeding ability have different economic functions within the herd, and their productive values will accordingly be differentially affected by exogenous shocks to the system This theory has been subjected to detailed empirical testing and found to be surprisingly well supported The equations for both disaggregated slaughter and slaughter weight behaved consistently with the theory[31]

---

[29] See Jarvis (1969, chap 8) for a discussion of this, including the importance of the labor force variable in the specification of the model

[30] See Jarvis (1969) for a discussion of the attempts to measure the effect of changes in the rate of inflation Nores (1972) has included in his slaughter equations the net change in the loan balance to the cattle sector to capture the effects of credit conditions The estimated coefficient on this variable is negative and significant in the slaughter equation for yearlings, cows, and bulls

[31] The relative beef/grain price is theoretically the primary determinant of both the optimal size of the herd and the optimal slaughter age of a given animal (the high price elasticity of consumer demand among different types of beef requires that relative prices

Second, the estimated equations yield solid evidence that cattle producers have responded in an economic manner to changes in the beef/grain relative price and to its rate of change The instantaneous response of slaughter to a price increase is negative for every animal category However, even though the lagged price coefficients (and their sum) are negative in most of the individual-category slaughter equations, the econometric model as a whole indicates that the long-run elasticity of slaughter is positive This fact has been frequently misunderstood The simple single-equation models estimated in most previous studies were essentially designed to show the negative short-run response of slaughter However, because these models cannot show that a rising herd, through higher calf crops, would lead to rising slaughter over time, they were often misinterpreted to imply a negative long-run elasticity as well This is incorrect The estimated price coefficients in my model (and most others) give only the effect of price on the "transitory" component of slaughter, not the "permanent" component A reduction of slaughter one year increases the size of the herd in the next (and therefore the permanent component of slaughter), and it is the net effect of changes in both the permanent and transitory components which yields the true effect of the price in lagged periods [32] Indeed, the greatest impact on future production comes from the increase in the calf crop, as females are withheld to increase the size of the breeding herd, and the equation estimating the number of calves born is crucial to the estimation of the long-run price elasticity [33]

I believe the results obtained should abolish doubts as to whether Argentina producers respond to prices These results, which have been obtained using herd data which I believe are much improved over those

---

of animals be rather constant, which in turn permits the price of one category to be used as a reasonable proxy for all in many econometric models) The climatic index, a proxy for certain types of feed costs, may also be important, but the interest rate plays a minor role in the slaughter decision because most animals are relatively short-lived Several measures of the sectoral terms of trade were included in preliminary estimations of the slaughter equations, but these variables were always insignificant This result may reflect more upon the poor quality of the variables used than upon the actual effect of the changing terms of trade on agricultural investment In nearly all similar studies it has been easier to obtain an intrasectoral supply response than an intersectoral response

[32] Yver (1971) has developed a method to determine the time profile of the net impact of exogenous disturbances on the dependent variables in a model similar (but necessarily more simple, for computational reasons) to the one presented here

[33] A discussion of the specification and estimation of this equation is contained in Jarvis (1969, chap 4, 1971) Models which estimate only the number of animals slaughtered, without considering changes in the average weight of the slaughtered animal, will also underestimate the long-run supply response because a price increase will cause animals within each category to be fed to somewhat heavier weights, and more animals will be withheld to older ages, e g , slaughtered as steers rather than yearlings This latter effect is frequently forgotten by those who note the relative constancy of individual-category slaughter weights, but it is a significant factor in total supply The cumulative elasticity of average dressed weight with respect to the beef/grain price is nearly 0 10

previously available, are statistically quite significant and identify the slaughter response of different animal categories The results show that producers systematically reallocate their portfolios in the expected manner when the recursive effect of their decisions is strongly evident, that is, with continuously operating markets for disposable productive assets And because much of the indicated response is an interactivity shift within the agricultural sector between grains and livestock, the price response shown by cattle producers implies a response by field crop producers as well [34]

It has been argued that discriminatory and capricious government price policy toward agriculture in the 1940s and 1950s, combined with rapid inflation and government intervention in land tenancy markets, so frustrated and confused producers in Argentina that they ceased to respond to the usual market indicators of profitability I find no support for this argument Although producers reacted to nonprice disturbances during the period studied, such as when government intervention in the tenancy market caused a reevaluation of the relative risk of different activities and thereby a demand shift for different productive factors, given this shift the sensitivity to market prices continued as before, that is, prices seem to have remained an important short-run determinant of production variation throughout the period

Third, the stronger is producer response, the more slaughter will be reduced in the short run This fact is important for a country like Argentina, which has frequently devalued its currency when the balance of payments has been in crisis The purpose of devaluation, which raises proportionately the domestic price of tradable goods (in the absence of export retentions, etc), is to reduce the consumption of exportables, like beef, while simultaneously stimulating internal production and increasing external demand for the same And the evidence suggests that devaluation in Argentina has usually increased beef exports in the first year because the decrease in domestic consumption and the increase in foreign demand outweigh the decrease in slaughter (as producers attempt to increase herds) However, because producer response leads to even higher prices, partially offsetting the effect of the previous devaluation in foreign markets and aggravating inflationary pressures in Argentina, beef exports frequently decline from their predevaluation levels in year 2 and even year 3. As a result, the government must be prepared to weather a period of almost 3 years before devaluation can significantly improve the balance of payments. The government generally has not been able to wait so long, and economic deflation has usually been induced to bring the balance of payments into temporary equilibrium.

---

[34] The asymmetry between these results and past studies of field crops may be due to measurement difficulties I have found, for example, that when corn production data are disaggregated at the *partido* (county) level, many individual *partidos* show significant price response even though their aggregate does not The degree of response for the individual *partidos* seems to be closely related to the level of the yields achieved and to the type of competing agricultural activities in the area One of the factors confounding the aggregate results is the difficulty of separating those corn plantings which are desired for grain harvest from those to be used for grazing cattle In some *partidos* a rise in the price of cattle is associated with a rise in the area planted to corn, but with a fall in the percentage of the planted area harvested (as corn is planted for forage), while in other *partidos* the beef price rise is associated with a decline in the area planted to corn

Price response clearly does not imply that producers allocate resources efficiently, even in the steady state. It suggests only that producers respond in the expected direction and that agricultural policies should take this fact into account. And the response encountered does not alter the fact that Argentine agriculture stagnated for nearly 3 decades. As price response is not the cause, per se, other explanations for this stagnation must be sought.[35]

References

Behrman, Jere. "Price Elasticity of the Marketed Surplus of a Subsistence Crop." *J. Farm Econ.*, vol. 49 (February 1967).

Colomé, Rinaldo A. "Funciones de oferta agropecuaria de la region Pampeana en el periodo 1940–1960." Mimeographed. Universidad Nacional de Cordoba, July 1966.

de Janvry, Alain. "A Socioeconomic Model of Induced Innovations for Argentine Agricultural Development." *Q.J.E.*, vol. 87, no. 3 (August 1973).

Diaz-Alejandro, Carlos F. *Exchange Rate Devaluation in a Semi-industrialized Country: The Experience of Argentina, 1955–1961.* Cambridge, Mass.: M.I.T. Press, 1965.

———. *Essays on the Economic History of the Argentine Republic.* New Haven, Conn.: Yale Univ. Press, 1970.

Freire, Remy. "Price Incentives in Argentine Agriculture." Mimeographed. Development Advisory Service, Harvard Univ., June 1966.

Hall, Robert E. "Polynomial Distributed Lags." Working Paper no. 7, Massachusetts Inst. Tech., Dept. Econ., July 1967.

---

[35] In my thesis (Jarvis 1969) I suggested that the principal cause of Argentine agricultural stagnation during the 1940s and 1950s was not lack of price response but the small amount of technical change which made its way into the Pampas. In response to changes in product and factor relative prices, producers shifted their resources among different economic activities and shifted substantially their use of factor inputs, largely by substituting cattle and machinery for migrating labor. Producers also adopted new techniques and inputs when these were available and profitable, including new cattle breeds, new seeds, and new machinery. Nonetheless, the total number of new innovations adopted was small relative to the experience of other countries possessing rich, well-developed agricultural sectors, and little increase in total factor productivity seems to have occurred in the Pampas. Government policy, which inhibited the introduction and dissemination of technical change during this period, must be held chiefly responsible. High tariffs and import prohibitions were imposed on many agricultural inputs, and the public agricultural research and extension system was effectively dismantled. Diaz-Alejandro (1970) also has suggested that government policies were most negative in their discouragement of technical change. However, de Janvry (1973) has argued that the large and politically influential agricultural producers have not encouraged government research in yield-increasing techniques, even in recent years, because these developments would not have been in their social and economic interest, and these producers have thus contributed to the lack of available technology.

Jarvis, Lovell S "Supply Response in the Cattle Industry The Argentine Case, 1937/38–1966/67 " Ph D dissertation, Massachusetts Inst Tech , 1969

———. "An Example of the Use of Economic Models to Construct Unavailable Data Series The Estimation of Disaggregated Cattle Herd Stocks for Argentina, 1937–1967 " Paper presented at Winter Meetings of Econometric Society, New Orleans, December 27–29, 1971 Published in Spanish as "Un ejemplo del uso de modelos económicos para la construcción de datos no disponibles la estimación de la existencia de vacuno desagregado en Argentina, 1937–1967 " *Económica* (Buenos Aires), no 1 (1973)

Nores, Gustavo "Quarterly Structure of the Argentine Beef Cattle Economy A Short Term Model, 1960–1970 " Ph D dissertation, Purdue Univ , 1972

Otrera, Wylian R. "An Econometric Model for Analyzing Argentine Beef Export Potentials " Ph D dissertation, Texas A & M Univ , 1966

Reca, Lucio G "The Price and Production Duality within Argentine Agriculture, 1923–1965 " Ph D dissertation, Univ Chicago, 1967

Williams, Hadley Cogswell "The Role of Agriculture in Argentine Development A Problem of Farmer Response " B A thesis, Harvard Univ , 1966

Yver, Raul E "The Investment Behavior and the Supply Response of the Cattle Industry in Argentina " Ph D dissertation, Univ Chicago, 1971

Rational Expectations and Bayesian Analysis

Richard M Cyert and Morris H DeGroot

Carnegie-Mellon University

The approach in this paper is the development of models that describe the process by which rational expectations may be developed within a market The concept of Bayesian learning is introduced Consistent and rational expectations are introduced in models where the firms cannot immediately move to equilibrium Three different models are developed which demonstrate the interaction of Bayesian learning and expectations in achieving a market equilibrium These models are dynamic and describe the transition process toward equilibrium Two of the models involve unknown parameters about which the firms learn The third is a control theory model explicitly involving adjustment costs

The concept of rational expectations was introduced in an important article by Muth (1961) Muth's hypothesis was that the mean expectation of firms with respect to some phenomenon, say price, was equal to the prediction that would be made by the relevant economic theory Muth gives his reasons for believing in the hypothesis, proceeds to develop it further, and does some testing as to the general consistency of the hypothesis Our objective in this paper is to build on Muth's basic concept by providing some insight into the processes by which the rational-expectations hypothesis can, in fact, be realized We will introduce the concept of Bayesian learning into expectations and then demonstrate for three different models how the concept might work in achieving a market equilibrium

This research was supported in part by the National Science Foundation under grant GS-32514 We are indebted to William Brock, David Cass, Michael Lovell, Robert Lucas, and Edward Prescott for helpful comments We would also like to thank an unknown referee for excellent comments that resulted in a significant improvement of the paper

The Process of Generating Rational Expectations

Muth sees his hypothesis developing out of a process in which each economic actor (he generally uses firms) has an expectation (prediction) about the value of some economic variable He makes it clear that each expectation is not perfect (in the sense of being accurate), nor are all the expectations the same Out of this process Muth hypothesizes that the weighted arithmetic mean of these expectations is equal to the prediction of the relevant economic model, generally the equilibrium value of the model It is our view that a process has to be developed if the rational hypothesis is to be a scientific truth rather than a religious belief

The attempt to develop process models immediately opens us to the criticism of developing ad hoc models We acknowledge that there may be a large number of models that could potentially describe the process to equilibrium Our position is that, while the models have a certain amount of face validity, our major contribution is the introduction of an explicit learning process described in Bayesian terms The notion of developing models to describe the transition process toward equilibrium of a system disturbed by some random shocks may be questioned by some economists The development of comparative statics and the neglect of dynamic analysis is in part a reflection of such attitudes in the profession Yet without well-developed process models, the concept of rational expectations is essentially a black box

Muth seems to assume an independence among the firms making the estimates Empirically, however, it is more reasonable to assume that the expectations are correlated The management of each firm in a given market operates in essentially the same kind of environment and has much of the same kind of information In most cases, the managers have also experienced the same set of historical inputs from having been in the particular industry for some time In addition, the existence of trade associations will tend to make similar information available to managers in the same industry, although managers can obviously obtain information from other sources It is also possible that even with similarity of information the same input can be interpreted somewhat differently when the subjective probability distributions are computed by each manager However, even if all firms do not initially have the same priors, the feedback from the market will tend to modify the priors to the extent that similarity becomes a reasonable assumption at some point Thus, we will postulate the same prior probability distributions for the decision makers in our models

Initially, Muth makes the assumption of independent shocks within his model He assumes that the error term is randomly distributed, with no serial correlation This model always leads to the trivial conclusion that the mean expectation is equal to the equilibrium value Thus, the existence of a cobweb is antithetical to Muth's notion, and in his paper he argues that the cobweb model is not a good description of the data As an alternative explanation, Muth has recourse to the fact that a dynamic system hit with random shocks exhibits cycles with a stable period Process models such as we develop may be a middle ground for explaining the phenomena They have neither to assume irrationality and a lack of learning, as is true of the traditional cobweb model, nor to rely solely on random processes for an explanation of the real world

Muth himself recognizes the fact that a model assuming rational expectations does not accommodate a dynamic process showing the path to equilibrium He attempts to accommodate such a process by introducing the notion of serially correlated disturbances He assumes that these disturbances can be represented as linear combinations of normally and independently distributed random variables and that both the actual price and the expected price can also be represented as linear combinations of these independent variables with appropriate weights This approach gives Muth's work greater flexibility However, in applications, the notion of independently and identically distributed disturbances is, for a variety of reasons, the most generally used As a result, most models incorporating rational expectations move directly to the equilibrium value of the model without specifying an adequate process to produce the result

Learning and Uncertainty

The model that we will follow will utilize Bayesian learning in describing the process by which expectations, once formed, lead to an equilibrium Rather than assume that market expectations lead directly to the equilibrium value, we will assume that learning takes place in the market and that this learning has the effect of continually modifying the prior probability distribution with which the firms start Through this process we will show that the market will typically converge to an equilibrium

The learning in our analysis comes from the feedback resulting from the actions taken by the firm If a firm has a prior distribution for a particular price and supplies the appropriate quantity, its prior will be modified and a new expectation formed when it observes the results in the market The new prior in turn leads to another decision The form of learning that takes place is learning about the unknown values of the parameters in the particular market in which the firm is operating The first step is the formation of a prior probability distribution about the values of the important parameters in the market The learning process then continues through feedback from the market and continuing modification of the prior (Cyert and DeGroot 1970, 1971)

We propose to deal with models in which there is uncertainty However, the term "uncertainty" takes on a completely different meaning when one moves from classical statistics to a Bayesian approach Uncertainty has been generally used to describe a situation in which the outcome of a decision maker's action is not precisely predictable because of the existence either of parameters with unknown values or of random error terms The traditional approach for analyzing uncertainty has been to assume a probability distribution for the error term but not for the unknown values of the parameters This procedure leads ultimately to the need for some choice criterion such as the minimax or Neyman-Pearson approach (Modigliani 1949, Arrow 1951, Savage 1954) When we move into the Bayesian world, these choice criteria no longer have relevance To the Bayesian, all uncertainty can be represented by probability distributions Thus, the approach is to assign a prior distribution to the unknown values of all the parameters wherever they appear in the model (DeGroot 1970) Learning then takes the form of changing conditional distributions for the outcomes in future periods on the basis of past outcomes The structure of the changes will vary depending on whether they arise from updating the posterior distributions of parameters or from serially correlated error terms The only conditions under which the learning process described in the preceding paragraph is not present are represented by models that consist of error terms independently distributed over time and no other parameters Our models will consist of both unknown values of parameters and independently distributed error terms The kind of learning we have described will be possible in these models In fact, as we shall now discuss, this kind of learning is possible even when the form of the model is unknown to the economic agents

An Inconsistent Model

We classify the first model that will be discussed as an inconsistent model because the firms do not know the form of the model of the process determining the price and, in fact, base their decision on an incorrect model Through learning from the feedback of market information, the firms can reach an equilibrium even though they have the wrong model However, since the model that the firm is using is incorrect, this process may converge only slowly or may not converge at all

We feel that the construction of inconsistent models is not the most fruitful way to progress in this field If the firms have models that diverge drastically from reality, it seems reasonable to assume that the management would recognize this condition and change the model The firms would continue searching for a model that produced predictions that coincide more closely with actual observations In fact, the term "rational expectations" coined by Muth actually describes models that are consistent The model under discussion, even though inconsistent, does have some interesting properties

This first model involves a parameter whose exact value is unknown to the firm The firm assigns a prior distribution to the value of the parameter, and after the price in each period is observed, the firm updates this distribution in accordance with Bayes's theorem However, since the price in each period is influenced by the firm's expectations, and these expectations are at least partly incorrect, the time series of prices that are generated could be highly unstable, even though other conditions of the model indicate a convergence to equilibrium should prevail

Assume that the firm's model for the price p_{t+1} in period $t + 1$, given the price p_t in period t, is of the following form

$$p_{t+1} = ap_t + v_{t+1}, \qquad t = 0, 1, 2, \qquad (1)$$

Here the value of the coefficient a is fixed throughout the process but is unknown, and v_1, v_2, form a sequence of independent and identically distributed random error terms Suppose also that the firm believes that the common distribution of each of the error terms is normal with mean 0 and known precision r (The precision of a normal distribution is the reciprocal of its variance)

If the firm has a given posterior distribution for the value of a after it has observed the price p_t, then this distribution becomes the new prior and the firm can apply Bayes's theorem to determine the posterior distribution of a after the price p_{t+1} has been observed In particular, it follows (DeGroot 1970, p 167) that if the posterior distribution of a at the end of a period t is normal with mean m_t and precision h_t, then the posterior distribution at the end of period $t + 1$ will be normal with mean m_{t+1} and precision h_{t+1}, where

$$m_{t+1} = \frac{h_t m_t + rp_t p_{t+1}}{h_t + rp_t^2} \qquad (2)$$

and

$$h_{t+1} = h_t + rp_t^2 \qquad (3)$$

It follows from these relations that at the end of period t, the firm's expectation $E_t(p_{t+1})$ of the price in period $t + 1$ will be

$$E_t(p_{t+1}) = E_t(a)p_t + E_t(v_{t+1}) = m_t p_t, \qquad (4)$$

since $E_t(a) = m_t$ and $E_t(v_{t+1}) = 0$

We shall assume that in fact, however, the actual price p_{t+1} depends on the firm's expectation $E_t(p_{t+1}) = m_t p_t$ Specifically, we shall assume that the market equations take the following form in any period t.

$$C_t = d_1 - \beta p_t \quad \text{(demand)}, \tag{5}$$

$$Q_t = d_2 + \gamma E_{t-1}(p_t) + u_t \quad \text{(supply)}, \tag{6}$$

$$Q_t = C_t \quad \text{(market equilibrium)} \tag{7}$$

In these relations, Q_t is the number of units produced in period t, C_t is the amount consumed in that period, and u_t is a random variable representing the uncontrollable factors which affect the amount produced in period t. It is assumed that d_1, d_2, β, and γ are given constants We have taken the same model as Muth to facilitate comparisons between his work and our own [1]

By solving the market equations, we obtain the relation

$$p_t = \frac{d_1 - d_2}{\beta} - \frac{\gamma}{\beta} E_{t-1}(p_t) - \frac{1}{\beta} u_t \tag{8}$$

We shall assume for simplicity, as did Muth, that $d_1 = d_2$ (so that the equilibrium price is 0), and we shall let $v_t = -(1/\beta)u_t$ Then we have

$$p_t = -\frac{\gamma}{\beta} E_{t-1}(p_t) + v_t \tag{9}$$

Thus, the price p_{t+1} is actually determined by the relation

$$p_{t+1} = -\frac{\gamma}{\beta} E_t(p_{t+1}) + v_{t+1}, \tag{10}$$

or equivalently

$$p_{t+1} = -\frac{\gamma}{\beta} m_t p_t + v_{t+1} \tag{11}$$

It is again assumed that $v_1, v_2,$ are independent and identically distributed, each having a normal distribution with mean 0 and precision r

Thus, although the firm believes that prices are determined in accordance with (1), and forms its expectations about the value of a in accordance with (2), prices are actually determined in accordance with (11) We have, therefore, a relatively sophisticated cobweb process As in other economic processes, the model becomes consistent at equilibrium Thus, if the firm becomes convinced that $a = 0$, so that $m_t = 0$, then the model is no longer inconsistent, because equations (1) and (11) become identical, p_{t+1} becoming equal to the error term v_{t+1}

---

[1] The existence of a random variable in the supply implies a random variable in the production function of the individual firms However, the existence of a random variable in the production function does not in any sense imply that the firms are not attempting to maximize profits It does mean that the firms must make their production decisions under uncertainty We are assuming that they do this in an optimum fashion, and equation (6) is the aggregate result of such decisions For a discussion of some related problems, see Grossman (1972)

In the conventional cobweb analysis, the attainment of equilibrium is dependent on the relative slopes of the demand and supply curves In terms of the parameters of our model, where the slope of the demand curve is $-(1/\beta)$ and the slope of the supply curve is $1/\gamma$,

$$\beta > \gamma \quad \text{convergence,} \tag{a}$$

$$\beta = \gamma \quad \text{oscillation,} \tag{b}$$

$$\beta < \gamma \quad \text{divergence} \tag{c}$$

These conditions do not hold for our model It can be seen from equation (11) that the condition for convergence is that $(\gamma/\beta)|m_t|$ ultimately becomes less than 1

Numerical Results

To get a feeling for the operation of the model, we made several Monte Carlo runs A given run is carried out by specifying the values of γ/β and r to be used throughout the run, as well as the initial price p_0, the mean m_0, and precision h_0 of the prior distribution of a, generating a sequence $v_1, v_2, \quad, v_n$ of independent and identically distributed errors for a fixed number of periods n and computing the corresponding sequence $p_1, p_2, \quad, p_n$ of prices

All of the runs were made using the value $h_0 = 0$ That is, it was assumed that the prior distribution of a was an improper, diffuse prior over the real line, although in order to determine the price p_1, it was still necessary to specify the value of m_0

As a first example, for $\gamma/\beta = r = p_0 = m_0 = 1$, the sequence given in table 1 was generated for a process lasting $n = 100$ periods As can be seen from table 1, the value of m_t is close to 0 after 100 periods, and the prices p_t, therefore, tend to differ little from the random error terms v_t It seems noteworthy, however, that even at $t = 100$, the values of m_t are still exhibiting small fluctuations In a second example, the initial price was changed to $p_0 = 10$, and the process was again followed for 100 periods using the same values of the random variates v_t The value of m_t fluctuated near $-0\ 03$ after 100 periods, and thus again p_t differed little from the random error terms

It is of particular interest that we are getting convergence in two cases that the conventional analysis classifies as oscillating This effect comes from the Bayesian learning built into the model The strength of the firms' beliefs in the correctness of their expectations can be seen in the growth of the precision h_t in table 1 (In the second example $h_t = 336$ after 100 periods) The large values of h indicate that the variance of the posterior

TABLE 1

INITIAL CONDITIONS $p_0 = m_0 = r = \gamma/\beta = 1$

t	m_t	h_t	p_t	$E_{t-1}(p_t)$	v_t
1	−0 539	1 000	−0 539	1 000	0 461
2	0 566	1 290	0 355	0 290	0 645
3	−0 198	1 416	1 267	−0 201	1 066
4	−0 043	3 021	0 119	−0 250	−0 132
5	−0 050	3 035	−0 177	−0 005	−0 182
96	0 084	76 458	0 101	0 031	0 132
97	0 085	76 468	0 695	0 008	0 703
98	0 085	76 951	0 063	0 059	0 122
99	0 083	76 955	−1 618	0 005	−1 612
100	0 054	79 572	1 309	−0 135	1 174

TABLE 2

INITIAL CONDITIONS $p_0 = m_0 = r = 1,\ \gamma/\beta = 0\ 1$

RUN	TEN RUNS OF 100 PERIODS			
	m_{100}	h_{100}	p_{100}	$E_{99}(p_{100})$
1	0 046	76 307	1 186	−0 119
2	0 030	105 143	−2 054	−0 010
3	−0 174	104 052	−0 598	−0 071
4	−0 094	97 175	2 160	−0 021
5	0 125	104 670	−1 969	−0 032
6	−0 004	100 390	−1 049	0 003
7	−0 005	92 691	−0 969	0 016
8	−0 081	80 874	−0 473	0 036
9	−0 017	95 349	−0 136	−0 010
10	−0 028	104 113	−0 397	0 001

distribution is very small The firms are beginning to treat their expectations as being based on essentially certain knowledge of the value of a

These results derive from single runs of the process and therefore do not necessarily represent average behavior of the process In table 2 the results of 10 independent runs for $p_0 = m_0 = r = 1$ are given, and in every run the model behaves as would be expected from conventional analysis

As the value of γ/β is increased beyond 1, the model produces an increased number of explosive solutions For $\gamma/\beta = 1\ 2$, we get convergence for all 10 runs For $\gamma/\beta = 1\ 5$, however, five of the 10 runs diverge in the sense that price p_t begins oscillating wildly, being positive for two periods, then negative for two periods, as its absolute value increases

Essentially the same kind of results prevails for different initial conditions of p_0 and m_0 For example, with $p_0 = 10$ and $m_0 = 1$, we get convergence of all 10 runs (100 iterations each) for $\gamma/\beta = 1\ 5$, in contrast

TABLE 3

INITIAL CONDITIONS $p_0 = m_0 = 10, r = \gamma/\beta = 1, v_t = 0$

t	m_t	h_t	p_t
1	-10	10^2	-10^2
2	9 802	$1\ 010 \times 10^4$	-10^3
3	$-9\ 606$	$1\ 010 \times 10^6$	$9\ 802 \times 10^3$
4	9 406	$9\ 709 \times 10^7$	$9\ 416 \times 10^4$
5	$-9\ 203$	$8\ 963 \times 10^9$	$-8\ 857 \times 10^5$
10	8 115	$3\ 085 \times 10^{19}$	$-4\ 601 \times 10^{10}$
15	$-6\ 873$	$2\ 810 \times 10^{28}$	$1\ 185 \times 10^{15}$
20	5 422	$4\ 334 \times 10^{36}$	$1\ 177 \times 10^{19}$
25	$-3\ 584$	$4\ 823 \times 10^{43}$	$-2\ 702 \times 10^{22}$
30	0 9246	$3\ 732 \times 10^{48}$	$-2\ 684 \times 10^{24}$
35	$-0\ 0773$	$1\ 761 \times 10^{49}$	$3\ 556 \times 10^{20}$
40	$-0\ 0773$	$1\ 761 \times 10^{49}$	$9\ 817 \times 10^{14}$
45	$-0\ 0773$	$1\ 761 \times 10^{49}$	$2\ 709 \times 10^9$
50	$-0\ 0773$	$1\ 761 \times 10^{49}$	$7\ 475 \times 10^3$
55	$-0\ 0773$	$1\ 761 \times 10^{49}$	0 0206
60	$-0\ 0773$	$1\ 761 \times 10^{49}$	0 0000

to the previous result for $p_0 = m_0 = 1$ But $\gamma/\beta = 1\ 75$ leads to all 10 runs exploding

An analytic study of the convergence or divergence of this process appears to be difficult In order to perceive the essential nature of the process when it is stripped of all stochastic elements, and to gain a better understanding of the Monte Carlo runs just described, calculations were carried out for a model in which the error term v_t in equation (11) was taken to be 0 in every period t

When $\gamma/\beta = 1$, it is found that the price p_t converges quickly to 0 for each of the following three sets of initial values (i) $p_0 = m_0 = 1$, (ii) $p_0 = 10$, $m_0 = 1$, (iii) $p_0 = m_0 = 2$ When $\gamma/\beta = 1$ and the initial values are $p_0 = m_0 = 10$, the results are as shown in table 3 This process appears to be exploding during the first 30 periods The price p_t spirals wildly outward, and again it is positive for two periods, then negative for two periods, as its absolute value increases The absolute value reaches a maximum of $2\ 684 \times 10^{24}$ at period $t = 30$ After that period, however, the price p_t moves rapidly toward 0 and the process converges

To assess the significance of the error term in this process, a Monte Carlo run was also carried out for these same initial values $p_0 = m_0 = 10$ and $\gamma/\beta = r = 1$, but with the error terms v_t again included in equation (11) Even with these error terms present, the process was found to behave similarly to the process represented in table 3 The price p_t reached its maximum absolute value of $1\ 60 \times 10^{24}$ in period $t = 29$ and then began to decrease in magnitude When $t = 29$, the values of m_t and h_t were found to be $m_t = 0\ 771$ and $h_t = 1\ 67 \times 10^{48}$ All these values are in approximate agreement with the values given in table 3

Our conclusion is that the expectational model brings the market to equilibrium under varying initial values, including some values when the constraints on the slope condition derived from conventional analysis indicate that divergence will occur It is clear that, if a mechanism were built into the model to allow the firm to modify its expectational models when they were clearly wrong, convergence could be attained with large ratios of γ/β and at a faster rate than we have observed in the current model Thus, one approach to further research of this kind (utilizing simulation) is to find learning mechanisms that will allow the economic units to change expectational models in a rational fashion

A Consistent Model

We will continue to work with the market equations described by equations (5)–(7), and we will now derive analytic solutions for a model that is free of any of the inconsistencies previously discussed We will assume that the values of some or all of the parameters d_1, d_2, β, and γ appearing in the demand and supply equations are unknown Moreover, the distribution of u_t may involve other parameters with unknown values

A consistent model means that the firm's expectations must be based on this mechanism Therefore, we can derive an explicit expression for these expectations by taking the expected value of both sides of equation (8) Thus,

$$\begin{aligned} E_{t-1}(p_t) &= E_{t-1}\left(\frac{d_1 - d_2}{\beta}\right) \\ &\quad - E_{t-1}\left(\frac{\gamma}{\beta}\right) E_{t-1}(p_t) - E_{t-1}\left(\frac{1}{\beta} u_t\right) \end{aligned} \qquad (12)$$

It then follows from this relation that

$$E_{t-1}(p_t) = \frac{E_{t-1}[(d_1 - d_2)/\beta] - E_{t-1}[(1/\beta)u_t]}{1 + E_{t-1}(\gamma/\beta)} \qquad (13)$$

Equation (13) represents the actual expectations of the firm in the market The market price will differ from this price because of the firm's incomplete knowledge of the values of the relevant parameters and random terms

The economic process will evolve as follows In each period, the firms form their expectation of the price in the next period from equation (13). The actual price is then generated in accordance with equation (8) The observed actual price leads the firms to change their expectations of the values of the parameters in equation (13) and, hence, to change their expectation of the price in the following period The next price is then generated from (8), and the process continues in this way.

As a simple example, suppose that in the market equations (5)-(7) the values of β and γ are known but that the value of $D = d_1 - d_2$ is not known with certainty. Suppose also that the random variables u_t are independent from period to period and that each u_t has a normal distribution with mean 0 and known precision τ (This assumption of known τ can be relaxed by the introduction of a prior distribution for τ when its value is unknown) It follows from (13) that

$$E_{t-1}(p_t) = \frac{E_{t-1}(D)}{\beta + \gamma} \tag{14}$$

Suppose now that the posterior distribution of D at the end of period $t - 1$ is normal with mean m_{t-1} and precision h_{t-1} Then $E_{t-1}(p_t) = m_{t-1}/(\beta + \gamma)$ Furthermore, equation (8) can be rewritten in the form

$$\beta p_t + \gamma E_{t-1}(p_t) = D - u_t \tag{15}$$

Therefore, it follows from Bayes's theorem that the posterior distribution of D at the end of period t will be a normal distribution with mean

$$m_t = \frac{h_{t-1}m_{t-1} + \tau[\beta p_t + \gamma E_{t-1}(p_t)]}{h_{t-1} + \tau}, \tag{16}$$

and precision $h_t = h_{t-1} + \tau$ Substituting (14) into (16), we get

$$m_t = \frac{\{h_{t-1} + [\tau\gamma/(\beta + \gamma)]\}m_{t-1} + \tau\beta p_t}{h_{t-1} + \tau} \tag{17}$$

It can be shown from these relations that

$$\operatorname*{plim}_{t\to\infty} m_t = D,$$

and, hence,

$$\operatorname*{plim}_{t\to\infty} E_{t-1}(p_t) = \frac{D}{\beta + \gamma} \tag{18}$$

Therefore, as $t \to \infty$, the expected price in each period becomes stabilized at the value $D/(\beta + \gamma)$ It follows from (8) that the distribution of the actual price in each period also becomes stabilized as a normal distribution with mean $D/(\beta + \gamma)$ and precision $\beta^2\tau$

Muth[2] distinguishes between two concepts, the prediction of the model (theory) $E_{t-1}(p_t)$ and the aggregate expectation of the firms p_t^e It is necessary for him to make the special assumption that $E_{t-1}(p_t) = p_t^e$, which he calls the hypothesis of rational expectations In our model and from the Bayesian point of view, this relation arises naturally and is nothing more than a necessary condition for the internal consistency of the

---

[2] Muth uses simply the notation $E(p_t)$, but it is clear that the meaning is the same

model we have presented here. In the above example, it is an immediate consequence of the hypothesis of rational expectations that $E_{t-1}(p_t) = p_t^e = D/(\beta + \gamma)$ In contrast, the consistent model we have presented shows the process leading to equation (18) If the firm is using a correct model, incorporating as many unknown parameters as might be needed, then equations (12) and (13) must be satisfied Indeed, it can be seen in equations (12) and (13) that it is in the Bayesian framework that the rational-expectations approach finds its natural setting

Sequences and Expectations

Up to this point we have dealt with the problem of rational expectations essentially in the way that Muth posed it The essential characteristic of that problem is that the firms are attempting to find the equilibrium price of output at every decision point Another, and more sophisticated, approach to the problem is to view the firms in the market as attempting to reach equilibrium in a specified number of stages or time periods The logic of moving in a sequence of stages stems from the fact that large and rapid changes in output are costly and, therefore, a smooth adjustment is sought Posing the problem in this way enables us to utilize control theory and, thereby, to determine the optimum output the firms should produce at each point These outputs then play the role of the equilibrium output in Muth's definition of rational expectations In this formulation, a rational sequence of expectations leads to the optimum sequence of outputs as determined from control theory

The Control Model[3]

In order to illustrate these ideas, we will work with the competitive model We will assume that all but one of the firms is operating at the long-run equilibrium price The lone firm not yet in equilibrium can be viewed as a new entrant or a firm that has been badly managed and has new management It is operating in a market with a fixed price, the long-run equilibrium price, and is attempting to maximize profits (minimize losses) in reaching the long-run equilibrium output

It is assumed that at the beginning of the process, the firm's output is at a certain given level Q_0 Then for $j = 1, \quad , n$,

$$Q_j = Q_{j-1} + u_j + e_j, \tag{19}$$

---

[3] The notation in this section is independent of that used in previous models in the paper

where u_j is the adjustment that the firm makes to its output Q_{j-1} when moving into period j, and e_j is a random factor affecting the output Q_j with mean 0 and variance σ_j^2. It is assumed that the random factors e_1, , e_n are independent If there is no random factor in any particular period, then $\sigma_j^2 = 0$ for that period

In a general formulation there would be a specific sequence of targets or goals Q_1^*, , Q_n^* with the property that the firm would like to have its output Q_j in period j close to the target Q_j^* The cost of being away from the target in period j is assumed to be $\alpha_j(Q_j - Q_j^*)^2$, and the cost of making an adjustment u_j in the output when moving into period j is assumed to be $\beta_j u_j^2$, where $\alpha_j \geq 0$ and $\beta_j \geq 0$ are given constants

In our particular example the targets Q_1^*, , Q_n^* are all equal to the same value Q^*, which is the long-run equilibrium output for this industry The cost $\alpha_j(Q_j - Q^*)^2$ reflects the distance of the firm's output in a given period from Q^* and includes the usual costs of production The cost $\beta_j u_j^2$ reflects such items as hiring and firing costs, overtime pay, supervision cost increases, and other such costs not included in the first term The symmetric quadratic form of these costs provides, under the usual differentiability conditions, an effective second-order approximation to a wide variety of actual cost functions

Since the two costs we have defined are separate, the total cost in a given period is the sum of the two, and the total cost of the process is the sum of all these costs for $j = 1$, , n Since we assume that the firm's utility function is linear, the problem is to choose the sequence of adjustments u_1, , u_n in order to maximize profits (minimize losses) Given the fixed market price, the firm achieves this objective by minimizing the overall expected cost

$$\sum_{j=1}^{n} E[\alpha_j(Q_j - Q^*)^2 + \beta_j u_j^2] \tag{20}$$

The optimal sequence of adjustments can be derived in this problem by the method of backward induction (DeGroot 1970, pp 411–14) and has the following properties The optimal adjustment u_j in period j is a linear function of the previous output Q_{j-1}, that is, $u_j = a_j Q_{j-1} + b_j$, where the constants a_j and b_j depend on Q^* and the values of α_j, , α_n and β_j, , β_n However, these optimal adjustments do not depend on the variances σ_1^2, , α_n^2

In the problem we are considering, we shall assume that the coefficients α_j and β_j which enter the cost function are of the form $\alpha_j = \alpha\rho^j$ and $\beta_j = \beta\rho^j$, where α and β are given positive constants and ρ is a discount factor $(0 < \rho < 1)$ Thus, the total expected cost of the process is

$$\sum_{j=1}^{n} \rho^j E[\alpha(Q_j - Q^*)^2 + \beta u_j^2] \tag{21}$$

It can be shown that if we let the number of periods $n \rightarrow \infty$, the optimal adjustment u_j converges to the following value, for $j = 1, 2,$

$$u_j = \gamma(Q^* - Q_{j-1}), \tag{22}$$

where γ is a certain constant in the interval $0 < \gamma < 1$ specified by the relation

$$\gamma = \frac{\sqrt{(\alpha + \beta - \rho\beta)^2 + 4\alpha\beta} + \alpha + \rho\beta - \beta}{\sqrt{(\alpha + \beta - \rho\beta)^2 + 4\alpha\beta} + \alpha + \rho\beta + \beta} \tag{23}$$

With this choice of the adjustment u_j, the output in period j will be generated by the model

$$Q_j = (1 - \gamma)Q_{j-1} + \gamma Q^* + e_j \tag{24}$$

Thus, with this derivation we can determine the optimal sequence of outputs for the firm to follow in adjusting to its ultimate target Q^* The values Q_1, , Q_n determine the least-cost path and help to determine another definition of rational expectations Thus, since the firm cannot reach the equilibrium value immediately, as is assumed in the Muth formulation, rational expectations in each period should be couched in terms of Q_1, , Q_n More specifically, if we return to the cobweb model which has occupied us in this paper, the firm's expectations about price in each period should lead to the appropriate output in the optimal sequence if the expectations are rational as defined

Formally, Q_j will be a function of $E_{j-1}(p_j)$, the firm's expectation of the price p_j in period j based on all observed prices and outputs up to the end of period $j - 1$ We shall let f denote this function, so that

$$Q_j = f[E_{j-1}(p_j)] + e_j, \tag{25}$$

where e_j again denotes a random factor with mean 0 and variance σ_j^2 which affects the output in period j

We now define the expectation $E_{j-1}(p_j)$ to be *rational* if this expectation yields a model for Q_j which agrees with the control theory model that was derived from the optimal sequence of adjustments In other words, $E_{j-1}(p_j)$ is a *rational expectation* if

$$f[E_{j-1}(p_j)] = (1 + a_j)Q_{j-1} + b_j \tag{26}$$

For the special problem with constant targets, discounted costs, and an infinite number of periods, $E_{j-1}(p_j)$ is a rational expectation if

$$f[E_{j-1}(p_j)] = (1 - \gamma)Q_{j-1} + \gamma Q^* \tag{27}$$

For example, suppose as in equation (6) that f is simply an increasing linear function of $E_{j-1}(p_j)$, as follows

$$f[E_{j-1}(p_j)] = c_1 E_{j-1}(p_j) + c_2 \tag{28}$$

Then $E_{j-1}(p_j)$ is a rational expectation if and only if

$$E_{j-1}(p_j) = \frac{1}{c_1}[(1-\gamma)Q_{j-1} + \gamma Q^* - c_2] \tag{29}$$

The sequence of expectations generated in this way has the property that the expectation is updated in each period j after the output Q_j in that period has been observed or, equivalently, after the price p_j in that period has been observed This updating is not "Bayesian" in the usual sense of the word, simply because there are no unknown parameters in this model about which the firm must learn Furthermore, the firm's expectations are not consistent or correct for the cobweb model in our earlier sense, because the actual expected price in period j, given the firm's choice of Q_j as determined by the market equations, will not be equal to $E_{j-1}(p_j)$ In what sense then are these expectations rational? They are rational in the sense that they lead the firm to an optimal output in each period, that is, they lead to a sequence of outputs that minimizes the total expected loss or, equivalently, maximizes the total expected profit

As a result of the application of control theory, it is possible to extend the definition of rational expectations to a process-oriented model There is no need to postulate a model in which the equilibrium price is attained by special assumptions about the average expectation Given the costs of moving to the equilibrium position, it is possible, as we have done, to determine the points on an optimum path A set of rational expectations should then lead the firm along this path The net effect will be to dampen the fluctuation of the cobweb model and carry the firms toward the equilibrium position

Conclusions

Our basic thrust in this paper has been the development of models that describe the process by which rational expectations may be developed within a market We accept the fundamental notion that phenomena based on nonrationality such as that exhibited in the cobweb model cannot exist for any period of significant duration However, we have also pointed out that no attention is given to the process by which rational expectations may develop in a market In particular, the learning process has been ignored and we have attempted to remedy this deficiency by introducing the concept of Bayesian learning We have also added the concept of consistent as well as rational expectations Finally, we have tackled the problem of expectations under conditions in which the firm cannot immediately move to an equilibrium position [4]

The work on expectations has its ultimate significance in its impact on the construction of dynamic models The development of dynamic models generally hinges on using some form of expectations as the means of enabling the decisions of one period to be made rationally More generally, expectations in a dynamic model become the basis for relating a current period to future periods We have indicated in this article some new approaches for incorporating expectations in dynamic models involving complex types of adjustments

References

Arrow, Kenneth J "Alternative Approaches to the Theory of Choice in Risk-taking Situations " *Econometrica* 19 (October 1951) 404–37

Cyert, R M , and DeGroot, M H "Bayesian Analysis and Duopoly Theory " *J P E* 78 (September/October 1970) 1168–84

——— "Interfirm Learning and the Kinked Demand Curve " *J Econ Theory* 3 (September 1971) 272–87

DeGroot, Morris H *Optimal Statistical Decisions* New York McGraw-Hill, 1970

Grossman, Sanford "A Bayesian Approach to Static, Stochastic Equilibria " Unpublished paper, H G B Alexander Res Found , Graduate School Bus , Univ Chicago, August 1972

Lucas, Robert E , Jr , and Prescott, Edward C "Investment under Uncertainty " *Econometrica* 39 (September 1971) 659–81

Modigliani, Franco "Liquidity and Uncertainty Discussion " *A E R , Papers and Proc* 39 (May 1949) 201–8

Muth, John F "Rational Expectations and the Theory of Price Movements " *Econometrica* 29 (July 1961) 315–35

Savage, L J *The Foundations of Statistics* New York Wiley, 1954

---

[4] It should be noted that we have retained Muth's approach of defining rational expectation in terms of conditions on just the mean value of a distribution and have attempted to expand the concept as described Lucas and Prescott (1971) have defined rational expectation of a variable to mean that the actual values and the anticipated values of the variable have the same probability distribution

Competitive Firm and the Theory of Input Demand under Price Uncertainty

Raveendra N Batra and Aman Ullah

Southern Methodist University

In this paper, we examine the behavior of the competitive firm faced with making input-hiring decisions under conditions of price uncertainty Unlike Sandmo and Baron, among others, we show categorically that a marginal increase in uncertainty stimulates a decline in the firm's output, provided the absolute risk aversion is decreasing Furthermore, the implications of a change in expected price remain unchanged, but those of a change in input price need not be the same as in the certainty case Specifically, the input demand function is downward sloping only if the production function is well behaved

The purpose of this paper is to develop systematically the theory of input demand for a competitive firm facing price uncertainty For the sake of simplicity, we postulate a two-input model and examine the implications of uncertainty for the firm's output and input demand under the assumption that the competitive firm, facing given input prices, seeks to maximize the expected utility from profits Earlier contributors to the theory of the firm have either approached the question from the side of product markets,[1] or, in examining the firm's behavior from the side of factor markets, they have assumed that the firm seeks to maximize expected profits and is thus risk neutral [2] The focus of this paper is on the

---

We are indebted to J Hadar and W Russell for useful discussions on the subject matter of this paper

[1] See, for example, McCall 1967, Horowitz 1970, chap 12, Sandmo 1971, and Leland 1972

[2] See, for example, Smith 1960, Tisdell 1968, chap 3, and Rothenberg and Smith 1971 One exception is that of Horowitz, who uses the expected utility approach to analyze the implications for the firm's decision making of uncertainty in the production function (not in the product price) in terms of a long-run model allowing for variability in the use of both capital and labor. Apart from the fact that Horowitz is concerned with different questions, his analysis is incomplete because he does not get involved in the comparative-statics properties of his model It may be mentioned in passing that in another chapter, Horowitz anticipates some of the results derived recently by Sandmo (1971), especially the results concerning the output level of the risk evader and the implications of a change in fixed costs

expected utility approach, which permits a more general formulation of the firm's attitude toward risk

For the main tools of analysis, we draw upon the recent seminal work by Sandmo (1971), who has examined under price uncertainty the behavior of the competitive firm from the side of the product market However, in contrast to Sandmo's results which are obtained from a short-run model of the firm where the capital stock is given, our results are derived from a long-run model where all factors of production are variable [3] Furthermore, Sandmo (1971) is unable to derive categorical effects of a marginal increase in uncertainty whereas we show that an increase in uncertainty leads to a decline in the firm's output, provided absolute risk aversion is decreasing

I. Assumptions and the Basic Model

Unless otherwise specified, the following assumptions will be maintained throughout the paper

1 The firm is a price taker in the probabilistic sense That is to say, the firm is not large enough to influence its subjective probability distribution about the sales price, which is assumed to be a random variable Furthermore, the prices of the two inputs, capital and labor, are given to the firm

2 The decisions concerning the volume of output and hence the utilization of inputs must be made prior to the knowledge of the market price

3 The firm seeks to maximize expected utility from profits, and its attitude toward risk can be described by a von Neumann–Morgenstern utility function, which among other things requires that the firm's decisions are made by one individual [4]

Let U stand for utility and π for profits Then the firm's utility function is given by

$$U = U(\pi) \tag{1}$$

where $U'(\pi) > 0$ and $U''(\pi) \lesseqgtr' 0$, depending on whether the firm is risk averse, risk neutral, or a risk preferrer [5] Let q = output, K = capital, and L = labor Then the production function of the firm is given by

---

[3] Actually, the short-run results of Sandmo can be derived as special cases of our results

[4] This assumption is central to much of the recent work in the theory of the firm under uncertainty (see, for example, McCall 1967, 1971, Sandmo 1971, and Leland 1972) Acceptance of this utility function implies acceptance of the axioms underlying the existence of the von Neumann–Morgenstern utility function For a lucid analysis of these axioms, see De Groot (1970) and Horowitz (1970, chap 12)

[5] Arrow (1965) has shown that if the utility function is to satisfy the von Neumann–Morgenstern axioms, it must be bounded from above This means that as π approaches infinity, the utility function must become concave, so that at least in the range close to the upper bound of the utility function, $U''(\pi) < 0$ This also suggests that for the firm to be risk neutral $[U''(\pi) = 0]$ or a risk preferrer $[U''(\pi) > 0]$, π should be restricted to a finite value

$$q = f(K, L) \tag{2}$$

The only assumption that we make about the production function is that f_K and f_L, the marginal products of capital and labor, respectively, are positive and that the isoquants are convex to the origin, so that $2f_K f_L f_{KL} - f_L^2 f_{KK} - f_K^2 f_{LL} > 0$, which, it may be noted, does not necessarily imply that $f_{ii} > 0$ $(i = L, K)$[6] The firm's total cost of production is given by $wL + rK$, so that its profit is given by

$$\pi = pq - wL - rK \tag{3}$$

where p is product price, which is assumed to be a random variable with density function $f(p)$ and mean $E(p) = \mu$ We assume that p is nonnegative

In view of (2) and (3), the expected utility from profit is furnished by

$$E\{U[pf(K, L) - wL - rK]\}, \tag{4}$$

where E is the expectations operator The firm chooses input quantities so as to maximize expected utility from profits The first-order conditions for the maximum are

$$\frac{\partial E(U)}{\partial L} = E[U'(\pi)(pf_L - w)] = 0 \tag{5}$$

and

$$\frac{\partial E(U)}{\partial K} = E[U'(\pi)(pf_K - r)] = 0 \tag{6}$$

The second-order conditions for the maximum are given by

$$\frac{\partial^2 E(U)}{\partial L^2} = A_1 = E[U''(\pi)(pf_L - w)^2 + pf_{LL}U'(\pi)] < 0, \tag{7}$$

$$\frac{\partial^2 E(U)}{\partial K^2} = A_2 = E[U''(\pi)(pf_K - r)^2 + pf_{KK}U'(\pi)] < 0, \tag{8}$$

and

$$\frac{\partial^2 E(U)}{\partial L^2}\frac{\partial^2 E(U)}{\partial K^2} - \left[\frac{\partial^2 E(U)}{\partial L\,\partial K}\right]^2 > 0, \tag{9*}$$

or

$$A_1 A_2 - B_1^2 = D > 0 \tag{9}$$

where

$$B_1 = \frac{\partial^2 E(U)}{\partial L\,\partial K} = E[U''(\pi)(pf_K - r)(pf_L - w) + pf_{KL}U'(\pi)] \tag{10}$$

---

[6] See, for example, Hadar (1970, chap 3, pp 26–27) on this point

Before we proceed further with our analysis, it is necessary first to briefly examine the second-order conditions for the attainment of maximum expected utility In the certainty case, utility maximization requires that $f_{ii} < 0$ and $f_{LL}f_{KK} - f_{KL}^2 > 0$ $(i = L, K)$, which means that the production function is strictly concave [7] We will now show that if the firm is risk averse, strict concavity of the production function is sufficient but not necessary to ensure expected utility maximization

Expanding (9*) by using (7), (8), and (10), we obtain

$$\begin{aligned} D = {} & \{E(U''a^2)E(U''b^2) - [E(U''ab)]^2\} \\ & + \{E(pU')[f_{KK}E(U''a^2) + f_{LL}E(U''b^2) - 2f_{KL}E(U''ab)]\} \\ & + \{[E(pU')]^2(f_{LL}f_{KK} - f_{KL}^2)\} \qquad (11) \end{aligned}$$

where $a = pf_L - w$, $b = pf_K - r$, and $U' \equiv U'(\pi)$ [8] From the first-order conditions given by (5) and (6), since f_L and f_K are nonrandom and since w and r are given,

$$\frac{w}{r} = \frac{f_L E[U'(\pi)p]}{f_K E[U'(\pi)p]} = \frac{f_L}{f_K} \qquad (12)$$

Using this in a and b, we get

$$a = (pf_L - w) = b\frac{f_L}{f_K} = \frac{f_L}{f_K}(pf_K - r), \qquad (13)$$

so that $a^2 = b^2 f_L^2 / f_K^2$ Using this, the first term in (11) reduces to zero, and D becomes $D = -E[U''b^2/f_K^2](2f_L f_K f_{KL} - f_L^2 f_{KK} - f_K^2 f_{LL}) + [E(pU')]^2(f_{LL}f_{KK} - f_{KL}^2)$ For second-order conditions to be satisfied, A_1 and A_2 from (7) and (8) should be each negative, whereas D should be positive If the firm is risk averse, then $U''(\pi) < 0$, in which case, A_1 and A_2 may be negative even if f_{KK} and f_{LL} are positive This immediately suggests that the production function need not be strictly concave for (7) and (8) to be satisfied This also holds true for (9), because D may be positive even if $(f_{LL}f_{KK} - f_{KL}^2) < 0$, provided, of course, that isoquants are convex to the origin so that $(2f_L f_K f_{KL} - f_L^2 f_{KK} - f_K^2 f_{LL}) > 0$ Thus, if the firm is risk averse, the concavity of the production function is no longer necessary for the firm to be at the optimum However, concavity is sufficient because it implies $f_{ii} < 0$ as well as convexity of the isoquants toward the origin

If the firm is risk neutral, then concavity is both necessary and sufficient for (7), (8), and (9) to be satisfied, because here $U''(\pi) = 0$ However, if the firm is a risk preferrer, so that $U''(\pi) > 0$, then the concavity of the production function becomes a necessary but not a sufficient condition for the firm to be at the optimum

---

[7] See, for instance, Henderson and Quandt 1971, pp 61–69

[8] Whenever the equation is too long, we will write U' for $U'(\pi)$

Suppose the firm is risk averse Then the convexity of the isoquants toward the origin along with (12) imply that, even with uncertainty about the product price, the firm minimizes the unit cost of production for whatever output it chooses to produce As with the certainty case, the marginal rate of substitution between the factors equals the factor price ratio at the optimum, even when the price is seen by the firm as a random variable This is very interesting, because in the following section we show that the risk-averse firm produces an output at which the expected value of the marginal product of each factor exceeds the given input price In other words, the firm produces an output at which expected price exceeds marginal cost, but the level of output is one where the average cost is minimum

II. Optimal Input Demand and Output under Uncertainty

In this section, we will show that under price uncertainty, the risk-averse competitive firm hires smaller quantities of inputs and thus produces lower output than in the case where p is known with certainty The problem may be formulated in terms of this question What are the optimal input demand and output under uncertainty as compared with the case where the product price is known with certainty to be equal to the mean of the probability distribution?

The first-order conditions (5) and (6) can be rewritten as

$$E[U'(\pi)pf_L] = E[U'(\pi)w] \tag{5*}$$

and

$$E[U'(\pi)pf_K] = E[U'(\pi)r] \tag{6*}$$

Subtracting $E[U'(\pi)\mu f_L]$ from both sides of (5*), we obtain

$$E[U'(\pi)f_L(p - \mu)] = E[U'(\pi)(w - \mu f_L)] \tag{14}$$

Applying the expectations operator to both sides of the profit equation (3) yields $E(\pi) = E(p)q - wL - rK$ or $E(\pi) - \mu q = -(wL + rK)$ Substituting the expression for $-(wL + rK)$ in (3) then furnishes $\pi = E(\pi) + (p - \mu)q$ If the utility function is strictly concave, so that $U''(\pi) < 0$, then $U'(\pi) \leq U'[E(\pi)]$ for $p \geq \mu$, or by multiplying through by $(p - \mu)f_L$, $U'(\pi)f_L(p - \mu) \leq U'[E(\pi)]f_L(p - \mu)$ for all p Taking expectations on both sides and noting that $U'[E(\pi)]$ is a given number and that f_L is nonrandom, we obtain

$$E[U'(\pi)f_L(p - \mu)] \leq 0 \tag{15}$$

Using this result in (14) implies that $E[U'(\pi)(w - \mu f_L)] \leq 0$, and since marginal utility is always positive, $w \leq \mu f_L$ Similarly, we can show that $r \leq \mu f_K$

In other words, under uncertainty equilibrium, the expected marginal value productivity of each factor exceeds its price, or, what is the same thing, $E(p)$ exceeds marginal cost By contrast, the certainty equilibrium is defined by the equality between marginal value productivity of each factor with its price Furthermore, under certainty conditions, $f_{ii} < 0$ is necessary to satisfy the second-order conditions for utility maximization Thus, if $f_{ii} < 0$, it is clear that under uncertainty, the optimal quantity demanded of each input is lower than the certainty case It then follows that the optimal output produced under uncertainty will also be lower than the certainty output

Until now, our analysis has been concerned with the implications for optimal input demand and output of what Sandmo calls the overall impact of uncertainty We will now consider the effects of a marginal increase in uncertainty by defining the increased variability of the density function of the price in terms of a "mean preserving spread"[9] Let us define p^* as $p^* = \gamma p + \theta$ where θ and γ are the shift parameters which initially equal zero and unity, respectively Then the mean preserving spread type of shift in the density function of p^* leaves the mean $E(p^*)$ unchanged, that is, $dE(p^*) = dE(\gamma p + \theta) = \mu\, d\gamma + d\theta = 0$, or

$$\frac{d\theta}{d\gamma} = -\mu \tag{16}$$

Using p^* instead of p, π and the first-order conditions are replaced by $\pi = (p\gamma + \theta)f(K, L) - wL - rK$, $E\{U'(\pi)[(p\gamma + \theta)f_L - w]\} = 0$, and $E\{U'(\pi)[(p\gamma + \theta)f_K - r]\} = 0$ Differentiating these conditions totally with respect to γ, and using (16), we obtain

$$A_1 \frac{\partial L}{\partial \gamma} + B_1 \frac{\partial K}{\partial \gamma} = H_1 \tag{17}$$

and

$$B_1 \frac{\partial L}{\partial \gamma} + A_2 \frac{\partial K}{\partial \gamma} = H_2 \tag{18}$$

where A_1, A_2, and B_1 have been defined before, and where

$$-H_1 = E[qU''(\pi)(pf_L - w)(p - \mu)] + E[U'(\pi)f_L(p - \mu)] \tag{19}$$

and

$$-H_2 = E[qU''(\pi)(pf_K - r)(p - \mu)] + E[U'(\pi)f_K(p - \mu)], \tag{20}$$

---

[9] Defining a change in uncertainty in terms of a change in the probability distribution while keeping its mean constant is quite common with economic theorists nowadays (see, for example, Sandmo 1970, 1971, 1972, Rothenberg and Smith 1970, and Rothschild and Stiglitz 1970, 1971)

remembering that initially $\gamma = 1$ and $\theta = 0$ Solving (17) and (18) simultaneously, noting that $H_1 = H_2 f_L/f_K$, and using (9), we obtain

$$\frac{\partial L}{\partial \gamma} = \frac{H_2[(A_2 f_L/f_K) - B_1]}{D} = \frac{H_2 E[U'(\pi)p](f_L f_{KK} - f_K f_{KL})}{D f_K} \qquad (21)$$

and

$$\frac{\partial K}{\partial \gamma} = \frac{H_2(A_1 - B_1 f_L/f_K)}{D} = \frac{H_2 E[U'(\pi)p](f_K f_{LL} - f_L f_{KL})}{D f_K} \qquad (22)$$

The signs of $\partial L/\partial \gamma$ and $\partial K/\partial \gamma$ depend on the signs of H_2 and f_{KL}, given the fact that $f_{ii} < 0$, and, from (9), D is positive Now the sign of H_2 from (20) depends on the signs of (i) $E[U'(\pi)(p - \mu)]$, which has already been shown to be negative under risk aversion, and (ii) $E[U''(\pi)(pf_K - r)(p - \mu)]$ In the next section, we show that under the hypotheses of decreasing absolute risk aversion, $E[U''(\pi)(pf_K - r)]$ is positive Thus

$$\begin{aligned} &E[U''(\pi)(pf_K - r)(p - \mu)] \\ &\quad = E\left\{U''(\pi)(pf_K - r)\left[\left(p - \frac{r}{f_K}\right) + \left(\frac{r}{f_K} - \mu\right)\right]\right\} \\ &\quad = E\left[U''(\pi)\,\frac{(pf_K - r)^2}{f_K}\right] + E[U''(\pi)(pf_K - r)]\left(\frac{r}{f_K} - \mu\right) \qquad (23) \end{aligned}$$

where $E\{U''(\pi)[(pf_K - r)^2/f_K]\} < 0$ if firms are risk averse and, as shown before, $(r/f_K) \leq \mu$ Therefore, with $E[U''(pf_K - r)]$ positive under decreasing absolute risk aversion, (23) is clearly negative All this discussion suggests that the right-hand side of (20) is negative, so that $H_2 > 0$

Once H_2 is demonstrated to be positive, then for $\partial L/\partial \gamma$ to be negative, it is clear that $(f_L f_{KK} - f_K f_{KL})$ is negative Similarly, for $\partial K/\partial \gamma < 0$, $(f_K f_{LL} - f_L f_{KL}) < 0$ Given that $f_{ii} < 0$, then a sufficient condition for the marginal impact of uncertainty to be in line with the overall impact is that $f_{KL} > 0$, which, of course, is satisfied for several well-known production functions[10] However, if $(f_L f_{KK} - f_K f_{KL})$ and $(f_K f_{LL} - f_L f_{KL})$ are positive because $f_{KL} < 0$, then the marginal impact of uncertainty will be opposite to the overall impact

However, although the marginal and overall effects of uncertainty on the firm's input utilization may not be the same, it still turns out that the marginal and overall impacts of uncertainty for the firm's output run in the same direction Differentiating the production function given by (2) with respect to γ and using (21) and (22), we obtain $(\partial q/\partial \gamma) = f_L(\partial L/\partial \gamma) + f_K(\partial K/\partial \gamma) = \{(2f_L f_K f_{KL} - f_L^2 f_{KK} - f_K^2 f_{LL})E[U'(\pi)p]H_2\}/Df_K$ It can be easily seen that with convex isoquants and $H_2 > 0$, $(\partial q/\partial y) < 0$ Thus, an increase in uncertainty leads to a decline in the firm's output, if absolute risk aversion is decreasing Note that Sandmo (1971), among many others, has been unable to obtain the predictable effects of a change in variance alone on the risk-averse firm's output

---

[10] A production function is said to be "well behaved" if $f_{ii} < 0$ and $f_{KL} > 0$ Thus, our results necessarily hold for well-behaved production functions

III. The Supply Function

This section is devoted to the analysis of the firm's supply function Specifically, we wish to determine the firm's output response to a change in the expected price The problem may be formulated as this Let price be written as $p + \theta$ where θ again is the shift parameter A change in θ shifts the probability distribution to the left or to the right and changes the expected value of price without altering the shape of the probability distribution

Differentiating (5) and (6) with respect to θ and evaluating the derivatives at $\theta = 0$, we obtain

$$A_1 \frac{\partial L}{\partial \theta} + B_1 \frac{\partial K}{\partial \theta} = -E[U''(\pi)(pf_L - w)q + f_L U'(\pi)] = F_1 \qquad (24)$$

and

$$B_1 \frac{\partial L}{\partial \theta} + A_2 \frac{\partial K}{\partial \theta} = -E[U''(\pi)(pf_K - r)q + f_K U'(\pi)] = F_2 \qquad (25)$$

The solution of this system yields

$$\frac{\partial L}{\partial \theta} = \frac{A_2 F_1 - B_1 F_2}{D} \qquad (26)$$

and

$$\frac{\partial K}{\partial \theta} = \frac{A_1 F_2 - B_1 F_1}{D} \qquad (27)$$

Since A_1 and A_2 are negative and $D > 0$, our problem now is to pin down the signs of B_1, F_1, and F_2 Using (13) in (10) gives us

$$\begin{aligned} B_1 &= E[U''(\pi)(pf_K - r)^2(f_L/f_K) + f_{KL}U'(\pi)p] \\ &= E[U''(\pi)(pf_L - w)^2(f_K/f_L) + f_{KL}U'(\pi)p] \end{aligned} \qquad (28)$$

Utilizing (12) and the new expression for B_1 and the expressions for F_1 and F_2 in (26) and (27), we obtain

$$\frac{\partial L}{\partial \theta} = \frac{E(U'p)\{E(U') + (q/f_K)E[U''(pf_K - r)]\}(f_K f_{KL} - f_L f_{KK})}{D}, \qquad (29)$$

$$\frac{\partial K}{\partial \theta} = \frac{E(U'p)\{E(U') + (q/f_L)E[U''(pf_L - w)]\}(f_L f_{KL} - f_K f_{LL})}{} \qquad (30)$$

We will show later in this section that $E[U''(pf_K - r)]$ and $E[U''(pf_L - w)]$ are positive under the hypothesis of decreasing absolute risk aversion or increasing absolute risk preference If we further assume that $f_{ii} < 0$ and $f_{KL} > 0$, it is clear that $\partial L/\partial\theta > 0$ and $\partial K/\partial\theta > 0$ These results continue to hold if the firm is risk neutral so that $U'' = 0$ In other words, a rise in the expected price will induce the firm to hire more inputs However, if $f_{KL} < 0$, then the results are not clear even if $f_{ii} < 0$ The expression for the change in output can be obtained by using (29) and (30) Thus, noting that $(pf_K - r)/f_K = (pf_L - w)/f_L$,

$$\frac{\partial q}{\partial \theta} = \frac{E(U'p)\{E(U') + (q/f_K)E[U''(pf_K - r)]\}\,(2f_L f_K f_{KL} - f_L^2 f_{KK} - f_K^2 f_{LL})}{D} \tag{31}$$

Evidently, risk neutrality, decreasing absolute risk aversion, or increasing absolute risk preference are sufficient to ensure that a rise in the expected price will induce the firm to increase its output It is interesting to note that the effects of an increase in price on input demand and output under price uncertainty are the same as the corresponding results available in the certainty case[11]

It is high time that we define the concepts of the firm's attitude toward risk Following Pratt (1964) and Arrow (1965), let us define the firm's risk attitude (R_a) as $R_a(\pi) = -[U''(\pi)]/[U'(\pi)]$, where $R_a(\pi)$ is commonly called the absolute risk-aversion function[12] For risk-averse firms, $R_a(\pi)$ is positive, for risk preferrers negative, and for risk-indifferent firms zero For risk-averse firms, Arrow suggests that $R_a(\pi)$ is decreasing in π This is the hypothesis of decreasing absolute risk aversion However, if $R_a(\pi) < 0$, it seems reasonable that $-R_a(\pi)$ be an increasing function of π If a firm prefers risk, then its risk preference should normally increase with the rise in profits This we call the hypothesis of increasing absolute risk preference

Following the line of reasoning suggested by Sandmo, we will now show that under these hypotheses $E[U''(\pi)(pf_K - r)] > 0$ Let $\bar{\pi}$ be the profit level when $p = r/f_K$ Since $(pf_K - r)/f_K = (pf_L - w)/f_L$, p also equals w/f_K If $R_a(\pi) > 0$ and $R_a'(\pi) < 0$, then $-[U''(\pi)]/[U'(\pi)] = R_a(\pi) \leq R_a(\bar{\pi})$ for $p \geq r/f_K$ where $R_a(\bar{\pi})$ is a given number Multiplying both sides by $-U'(\pi)(pf_K - r)$, we obtain $U''(\pi)(pf_K - r) \geq$

---

[11] In the certainty case, it can be easily seen that $\partial L/\partial p = [p(f_K f_{KL} - f_L f_{KK})]/D$, $\partial K/\partial p = [p(f_L f_{KL} - f_K f_{LL})]/D$, and $(\partial q/\partial p) = [p(2f_L f_K f_{KL} - f_L^2 f_{KK} - f_K^2 f_{LL})]/D$ The effects of a change in p on the input demand and output under certainty as well as uncertainty are the same in the sense that, in both situations, the effects are determined by the signs of the same expressions Now, if the same expressions have different signs under certainty versus uncertainty, then that is a different matter However, if we assume concave production functions, then the effects of price changes in both situations are similar

[12] Also, see McCall 1971 and Sandmo 1971

$R_a(\bar{\pi})U'(\pi)(pf_K - r)$ for all p Taking expectations on both sides and noting (6), $E[U''(\pi)(pf_K - r)] \geq R_a(\bar{\pi})E[U'(\pi)(pf_K - r)] = 0$

The same proof also applies to the case where $-R_a > 0$ and $-R_a'(\pi) > 0$ Here $[U''(\pi)]/[U'(\pi)] = -R_a(\pi) \geq -R_a(\bar{\pi})$ for $p \geq r/f_K$, so that $E[U''(\pi)(pf_K - r)] \geq -R_a(\bar{\pi})E[U'(\pi)(pf_K - r)] = 0$ for all p Therefore, $E[U''(\pi)(pf_K - r)]$ is positive

IV. The Input Demand Functions

We now turn to the effect of a change in the price of any input on the firm's demand for the two factors For the sake of simplicity, suppose there occurs a change in the wage rate but r remains unaltered Differentiating the first-order conditions (5) and (6) totally with respect to w alone and remembering that $dr = 0$, we obtain $A_1(\partial L/\partial w) + B_1(\partial K/\partial w) = C_1$ and $B_1(\partial L/\partial w) + A_2(\partial K/\partial w) = C_2$ where $C_1 = LE[U''(\pi)(pf_L - w)] + E[U'(\pi)]$ and $C_2 = LE[U''(\pi)(pf_K - r)]$ Following exactly the procedure established in the earlier section, we find that

$$\frac{\partial L}{\partial w} = \frac{A_2E(U')f_K + LE(U'p)E[U''(pf_K - r)](f_L f_{KK} - f_K f_{KL})}{Df_L},$$

$$\frac{\partial K}{\partial w} = -\frac{A_1E(U')f_K + [E(U'p)]\{LE[U''(pf_L - w)] + E(U')\}(f_K f_{LL} - f_L f_{KL})}{Df_L}, \tag{32}$$

and

$$\frac{\partial q}{\partial w} = \frac{1}{Df_L}\{LE[U''(pf_K - r)](f_L^2 f_{KK} + f_K^2 f_{LL} - 2f_L f_K f_{KL}) + f_K E(U')(f_L f_{KK} - f_K f_{KL})E(U'p)\} \tag{33}$$

The results, as usual, depend on our hypotheses of the firm's attitude toward risk as well as the signs of f_{KL} and f_{ii} In the certainty case, $\partial L/\partial w < 0$, whereas the signs of $\partial K/\partial w$ and $\partial q/\partial w$ depend on the sign of f_{KL}[13] These results continue to hold when the firm is risk neutral for then $\partial L/\partial w < 0$, whereas the sign of $\partial q/\partial w$ depends, among other things, on the sign of f_{KL} In the case of $\partial K/\partial w$, once it is realized that with risk neutrality A_1 now reduces to $f_{LL}E[U'(\pi)p]$, it can be clearly seen that the sign of $\partial K/\partial w$ depends only on the sign of f_{KL}

The results, however, are not so clear and interpretable when the firm is not risk indifferent If we assume that $f_{KL} > 0$ and $f_{ii} < 0$, then under the hypothesis of decreasing absolute risk aversion or increasing absolute risk preference which ensures that $E[U''(\pi)(pf_K - r)] > 0$, $\partial L/\partial w < 0$ and $\partial q/\partial w < 0$ However, the sign of $\partial K/\partial w$ is still indeterminate because the second term in (32) is negative, whereas the first term, with $A_1 < 0$ from (7), is positive

---

[13] In the certainty case, $\partial L/\partial w = (pf_{LL})/(f_{LL}f_{KK} - f_{KL}^2)$, $\partial K/\partial w = -(pf_{KL})/(f_{LL}f_{KK} - f_{KL}^2)$, and $\partial q/\partial w = [p(f_L f_{LL} - f_K f_{KL})]/(f_{LL}f_{KK} - f_{KL}^2)$

The differences between these results and those available under certainty must now be ostensible The main differences are

i) Under uncertainty, the sign of f_{KL} also enters into the determination of the sign of $\partial L/\partial w$, whereas in the certainty case f_{KL} plays no such role, and as a result the sign of $\partial L/\partial w$ is unambiguously negative

ii) Under certainty, the sign of $\partial K/\partial w$ is determined solely by the sign of f_{KL}, whereas under uncertainty the sign of $\partial K/\partial w$ is ambiguous

The crux of this discussion is that with well-behaved production functions (so that $f_{ii} < 0$ and $f_{KL} > 0$) and with our hypotheses concerning the firm's risk attitude, a rise in the price of any input induces the firm to reduce the demand for that input and lower its output However, the firm's response toward the demand for the other input is indeterminate

V Conclusions

Using a two-input, long-run equilibrium model of the competitive firm facing price uncertainty, we have derived the following results in the foregoing sections

1 The concavity of the production function is both necessary and sufficient to ensure the firm to reach an optimum only if the firm is risk neutral If the firm is risk averse, concavity is sufficient but no longer necessary, whereas if the firm is a risk preferrer, the opposite is true, provided the firm's objective is to maximize expected utility from profits

2 The risk-averse firm utilizes smaller quantities of inputs and thus produces a lower output than a firm operating under certainty, even though the former minimizes the unit cost of production of whatever output it chooses to produce Furthermore, an increase in uncertainty leads to a decline in the firm's output if its absolute risk aversion is decreasing in profits

3 The marginal and overall impacts of uncertainty need not be the same for input demand, although for the output level, they continue to be the same

4 The presence of uncertainty makes no difference to the effects of an increase in price on input demand and the output

5 In the certainty case, a rise in the price of an input induces the firm to lower the utilization of that input In the uncertainty case, however, we need a well-behaved production function to ensure this result Furthermore, under uncertainty the effects on the demand for the other input are unpredictable

References

Arrow, K J *Aspects of the Theory of Risk Bearing* Helsinki Jahnssonin Säätiö, 1965

De Groot, M H *Optimal Statistical Decisions.* New York McGraw-Hill, 1970

Hadar, J *Mathematical Theory of Economic Behavior* Reading, Mass · Addison-Wesley, 1971

Henderson, H , and Quandt, R E *Microeconomic Theory—a Mathematical Approach* 2d ed New York McGraw-Hill, 1971

Horowitz, I *Decision Making and Theory of the Firm* New York Holt, Rinehart & Winston, 1970

Leland, H E "Theory of the Firm Facing Uncertain Demand " *A E R* (June 1972), pp. 278–91

McCall, J J "Competitive Production for Constant Risk Utility Functions " *Rev Econ Studies* 34 (October 1967) 417–20

——— "Probabilistic Microeconomics " *Bell J Econ and Management Sci* (Autumn 1971), pp 403–33

Pratt, J W "Risk Aversion in the Small and in the Large." *Econometrica* 32 (January/April 1964) 122–36

Rothenberg, T J , and Smith, K R "The Effect of Uncertainty on Resource Allocation in a General Equilibrium Model " *Q J E* 85 (August 1971) 440–59

Rothschild, M , and Stiglitz, J E "Increasing Risk I A Definition " *J Econ Theory* 2 (September 1970) 225–43

——— "Increasing Risk II Its Economic Consequences " *J Econ Theory* 3 (March 1971) 66–84

Sandmo, A "The Effect of Uncertainty on Saving Decisions " *Rev Econ Studies* 37 (July 1970) 353–60

——— "Competitive Firm under Price Uncertainty " *A E R* (March 1971), pp 65–73

——— "Discount Rates for Public Investment under Uncertainty " *Internat Econ Rev* 13 (June 1972) 287–302

Smith, K R "The Effect of Uncertainty on Monopoly Price, Capital Stock and Utilization of Capital " *J Econ Theory* 1 (June 1970) 48–59

Tisdell, C A *The Theory of Price Uncertainty, Production and Profit* Princeton, N J Princeton Univ Press, 1968

Lifetime Earnings in a Professional Labor Market: Academic Economists

---

George E Johnson and Frank P Stafford

University of Michigan

It is hypothesized that persons who are more able and who enter the profession earlier in life purchase, through foregone earnings, more training options upon receipt of the Ph D and hence have more precipitous earnings profiles than those who are less able and who enter later By using National Science Foundation Register data from 1964 to 1970, earnings of economists and other academics are compared by age cohort both cross-sectionally and over time Defining junior and senior faculty as the two inputs producing instructional services and assuming a homothetic production function, a substitution elasticity of about five is estimated

I. Introduction

Analysis of the determinants of the level and structure of academic earnings is interesting both in its own right and for what it reveals about the behavior of professional labor markets This paper focuses principally on the question of differences in the pattern of earnings over the work cycle of academics In particular, we are interested in the way certain indirect measures of "quality" and of the time path of investments in human capital influence earnings at various levels of experience The empirical analysis of this paper is primarily limited to academic economists, but most of our major findings have been verified in other disciplines The data sources are the 1964 and 1970 National Registers

This research was sponsored by the National Science Foundation (GS-31639) and in cooperation with the American Association of University Professors We would like to thank Ronald Oaxaca, Sherwin Rosen, T W Schultz, and P O Steiner for comments on earlier versions of the paper Opinions expressed in this paper are the authors' We would like to thank William Hershey and Robert Gitter for providing able assistance in this project Operation with the large and varied data sets used in this paper was greatly facilitated by use of a computer software language, Micro, developed by the Labor Market Information Systems Project, directed by Malcolm Cohen

collected by the National Science Foundation (NSF),[1] and while the Register data have already been used to analyze the earnings of economists (Tolles and Melichar 1968), the orientation of their research was much broader in coverage and did not focus on the questions with which we deal.

The time period of our analysis corresponds to a slowing down in the growth rate of faculty salaries as has been demonstrated by American Association of University Professors (AAUP) studies.[2] Economists experienced a similar though less pronounced downturn toward the latter part of the period, with average price-adjusted growth rates by rank falling off from over 3 percent in 1965–66 and 1966–67 to only minor growth in 1968–71 (Boddy 1971). Moreover, to the extent that rates of promotion also slowed down, the relation between years of experience and earnings would have been lessened to a greater degree than suggested by rank-specific salary changes. The impact of this slowed rate of growth of salaries will be reflected in our 1970 data but will probably have a greater effect on the structure of earnings at some later date. Yet there were substantial changes in the structure of earnings which were already present in mathematics and physics by 1970.

The plan of the paper is as follows. In Section II we develop the basic model of the determinants of lifetime academic salaries and estimate its parameters on the 1970 data set. In Section III, we add an indirect measure of quality differences among graduates—whether or not the respondent attended a highly ranked graduate school—and estimate its effect on academic salary at various stages of the work cycle. Section IV characterizes lifetime earnings of economists in different types of institutions. Section V presents comparisons of earnings between economists and practitioners of a selection of other disciplines, and Section VI offers a summary of our major findings and their implications.

II. Effect of Professional Experience on the Salaries of Academic Economists

Although it is widely recognized that actual labor-market experience is a much more important determinant of earnings capacity than age and is often more important than formal schooling, most microdata sets contain age but not years of actual experience in present occupation. This requires that years of potential labor-force experience, defined as

---

[1] We would like to thank James Brown of the National Science Foundation for his support in providing these data from the National Register Files.

[2] See American Association of University Professors (1972). There it is shown that the real rate of growth of all faculty salaries declined as follows: 1965–67, 3.2 percent; 1967–68, 1.4 percent; 1968–69, 1.1 percent; 1969–70, −0.6 percent; 1970–71, −1.2 percent.

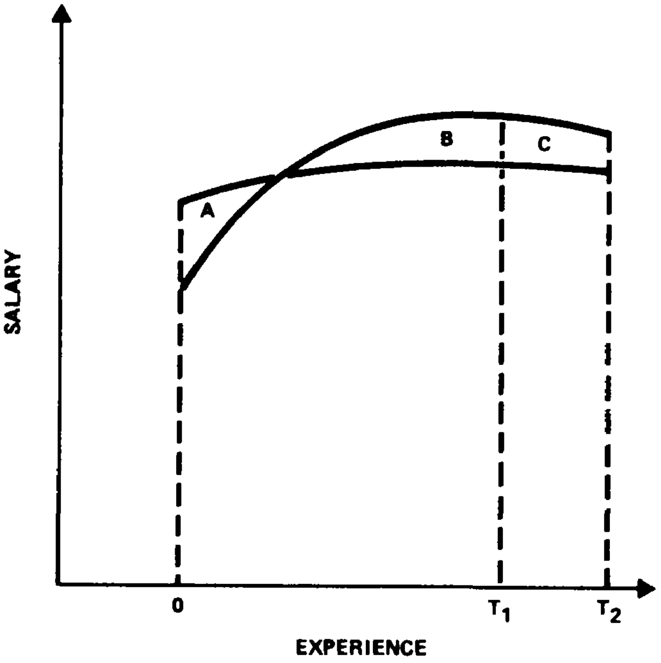


FIG 1 —T_1 = years worked by late starter T_2 = years worked by early starter A = earnings of late starter, $B + C$ = earnings of early starter

age minus estimated age at graduation from school, be employed as a proxy for actual experience For most professional occupations this is not an entirely satisfactory procedure For a variety of reasons (e g , military service, childbearing, indecision) individuals get started on their professional careers at many different ages, and the skills acquired in past pursuits are usually not transferable to their professions The NSF data do include the respondent's report of years of professional experience in present occupation, and this will allow us to estimate directly, and possibly more accurately, the relation between earnings and experience

There are theoretical reasons for expecting the observed relation between earnings and experience to vary systematically with the age at which individuals begin their professional experience First one way in which a professional forms human capital is to start out in a job in which he can learn a great deal about his trade Jobs with a relatively low training component must pay relatively high salaries, for an individual who takes one of these is presumably sacrificing future earnings (Mincer 1962, Rosen 1972) A late starter in a particular occupation will have less time than an early starter to capture the future gains from holding a job with a high training component and thus will be less likely to take such a job. This argument is summarized geometrically in figure 1. A late starter expects to work T_1 years, the early starter T_2 years. While the late starter is willing to sacrifice the future earnings represented by B in order to get A, the early starter finds it to his advantage to sacrifice A in order to get B plus C in the future. In other words, the late starter has less time remaining to make use of acquired skills, and thus the discounted present value of marginal benefits is lower (Ben-Porath 1967; Stephan 1971).

Utilizing the years of reported-experience variable, we were able to infer that individuals' definitions of "years of professional experience" varied considerably. For example, if we examine the relation between years of professional experience and age at which highest degree was attained, we find that some people claimed many more years of professional experience than elapsed years since Ph.D. (highest degree). Now for some individuals, this difference may truly reflect professional experience, for example, a year or two spent completing the finishing touches on a dissertation while fully employed. On the other hand, for others these years probably reflect little more than time since receipt of a B.A.—any labor-force experience no matter how remotely pertinent to current profession. To the extent that the latter is true, this predegree experience would result in their being the "late starters" portrayed in figure 1. Whether or not the predegree experience reflects primarily elapsed time or true professional experience is thus an empirical question. To proceed in our investigation, we first estimated a simple quadratic in years of reported experience for academic economists who have received the doctorate and worked full time.[3] We estimated the parameters of the following basic model: $\ln SAL = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 FEM + \varepsilon$, where SAL = 9–10 month academic salary, X = years of reported professional experience, and FEM = 1 if female, 0, otherwise.

From previous research the expectation is that $\alpha_1 > 0$ and both α_2 and $\alpha_3 > 0$.[4] The empirical results for 3,040 academic economists with this specification are given as

$$\begin{array}{llllll} \ln SAL = & .0333X & - .000518X^2 & - .153FEM & + 4.727 & \\ & (.0013) & (.000033) & (.018) & (.0097) & (1) \end{array}$$

$$R^2 = .363, \text{ standard error of estimate} = .209.$$

To test the hypothesis that much of the experience of some respondents is predegree activity of nominal value in current occupation, we dis-

---

[3] For an extensive justification of the use of a quadratic in experience to explain the logarithm of earnings, see Mincer (1973).

[4] The salary of females relative to males is thus $SAL_F/SAL_M = e^{\alpha_3}$, which is less than one if $\alpha_3 < 0$, and the proportionate-wage-disadvantage of females is

$$\frac{SAL_M - SAL_F}{SAL_M} = 1 - e^{\alpha_3}.$$

aggregated total reported experience through use of the year-of-Ph D variable The model then was

$$\ln SAL = \beta_0 + \beta_1 XPO + \beta_2 (XPO)^2 + \beta_3 XPR + \beta_4 (XPR)^2$$
$$+ \beta_5 XPO \times XPR + \beta_6 FEM + \varepsilon,$$

where XPO = postdegree experience and XPR = predegree experience ($X = XPO + XPR$) Our expectation is that β_1, $\beta_3 > 0$, $\beta_1 > \beta_3$, β_2, $\beta_4 < 0$, $\beta_5 < 0$, and that $\beta_1 + 2\beta_2 > \beta_3 + 2\beta_4$ over the range of 0–10 years of XPO and XPR (89 percent have XPR of 10 years or less) The results for the specification with reported experience disaggregated as we have suggested are

$$\ln SAL = \underset{(0013)}{0441XPO} - \underset{(000036)}{000771(XPO)^2} + \underset{(0021)}{0098XPR} - \underset{(000091)}{000182(XPR)^2}$$
$$- \underset{(000100)}{000656XPO \times XPR} - \underset{(016)}{139FEM} + \underset{(009)}{4\,735} \qquad (2)$$

$$R^2 = 457, \text{ standard error of estimate} = 193$$

The test of the null hypothesis that disaggregation of total experience into predegree and postdegree experience does not add to the explanation of salaries is decisively refuted ($F = 175\,0$), and we employ this functional form throughout the remainder of the paper[5]

The results of equation (2) show that the rate of salary increase is more than four times larger for postdegree experience than for predegree experience (0441 in comparison to 0098) at 0 years of each type of experience The substantial (negative) interaction terms between years of post- and predegree experience imply that the longer the period of predegree experience, the flatter the postdegree experience profile This result is fully consistent with our early starter–late starter distinction represented in figure 1[6] In addition to disaggregated reported experience, we also introduced a variable representing age at receipt of Ph D While

---

[5] Since the different profile shapes for early and late starters are critical, we also disaggregated our sample into five XPR categories 0, 1–2, 3–5, 6–10, 11+ This test provided the same information about profile shapes as eq (2), and a test on this disaggregation yielded $F = 162\,1$ The basic result in eq (2) obtains in all other disciplines we studied physics, anthropology, mathematics, sociology, and biology

[6] Equation (2) was estimated for a subgroup of professional economists for whom predegree experience would be expected to be more important, namely, business economists (These were excluded from our previous analysis as mentioned in n 9) The corresponding coefficients of 0407 and 0154 on XPR for this group compared to 0441 and 0098 in eq (2) give some additional support to the notion that pre- and postdegree experience represent different rates of professional development in different academic areas It should be noted at this point that the coefficient on FEM declines when experience is disaggregated because women have more years of predegree experience Eor a more complete discussion of this and related issues, see Johnson and Stafford (1974)

TABLE 1

PREDICTED SALARIES OF ACADEMIC ECONOMISTS BY YEARS OF POST-PH.D. EXPERIENCE FOR SELECTED LEVELS OF PRE-PH.D. EXPERIENCE

	XPO							
XPR	0	5	10	15	20	25	30	35
0	11,390	13,920	16,380	18,550	20,210	21,180	21,360	20,730
5	11,910	14,330	16,580	18,471	19,790	20,410	20,250	19,330
10	12,350	14,610	16,640	18,230	19,214	19,490	19,020	17,860
SAL (0)	0.922	0.953	0.985	1.018	1.052	1.087	1.123	1.160
SAL (10)	(.013)	(.011)	(.010)	(.011)	(.014)	(.019)	(.025)	(.031)

NOTE.—Standard errors of the estimated values of log *SAL*(0) − log *SAL*(10) are given in parentheses beneath the estimated values of *SAL*(0)/*SAL*(10). These are calculated according to the formula

$$s = 10[V(\hat{\beta}_3) + 100V(\hat{\beta}_4) + XPO^2V(\hat{\beta}_5) + 20C(\hat{\beta}_3, \hat{\beta}_4) + 2XPOC(\hat{\beta}_3, \hat{\beta}_5) + 20XPO^2C(\hat{\beta}_4, \hat{\beta}_5)]^{1/2}$$

the estimated coefficient on this variable was on the margin of statistical significance, it was of small quantitative importance and did not hold in other subsamples in our analysis. Accordingly, we adhere to our simpler specification in equation (2).

The economic significance of the results in equation (2) can be inferred by noting that the mean years of predegree experience is 4.3 years in our sample of full-time Ph.D. academics and has a standard deviation of 4.8 years.[7] Using these figures as the basis for approximate comparisons in table 1, we have presented profiles for individuals with different years of predegree experience at 5-year intervals of postdegree experience. A person with *XPR* = 0 would be typically someone who went straight through graduate school and began at about age 25, whereas a person with *XPR* = 10 would have entered the profession at, say, age 35 with a Ph.D. The observed starting-salary differential in favor of the late starter of $900 per year can be interpreted as reflecting his choice of lower training options relative to the early starter, however, the accumulated training differentials eventually result in the early starter's overtaking the late starter (at *XPO* = 10–15) and surpassing him in earnings from there on.

The empirical importance of disaggregating experience is summarized in table 1. Yet the question of interpretation of the difference still remains open. This is because the age of entry implied in *XPR* probably relates to ability as well as to the time period over which benefits are expected to accrue. One extreme would be to interpret the entire difference in profile shapes to differences in ability. There is empirical evidence that the more able students, at least by conventional measures, are those who receive sufficient stipend support to allow them to study full time rather than enroll part time and hold outside employment which prolongs their stay in graduate school (see Warkov 1965, Mooney 1968) On the other hand, some students spend extra time completing a thesis which is expected to have a very high payoff While there probably are, on net, differences in ability in favor of early starters, a tentative interpretation is that the shapes reflect primarily differences in remaining work life and associated differences in the expected span of remaining labor-force years In this regard, it is interesting to see if for the estimated relation in (2) the implied rate of return from choosing more or less on-the-job training is in line with plausible discount rates Suppose the individuals' abilities were equally distributed across the implicit age-of-entry variable (*XPR*) and that arbitage took place across the market for training options Then, under a discounted lifetime income–maximizing hypothesis, we would expect the rate of return to choosing the training option path (described as "early starters" in fig 1) to equal some reasonable discount rate such as the rate of return on alternative investments—say, the long-run rate of return on physical capital, which can be given as 8 percent from studies of the rate of return in manufacturing (Stigler 1964) From equation (2) the (marginal) rate of return to choosing the training option for someone entering the profession at age 30 (after 5 years of predegree experience) can be calculated as 7 0 percent [8] Given an assumed upward shift of 2 percent per annum in both earnings profiles (owing to generalized productivity growth and ignoring rising marginal taxes), the implied rate of return to choosing training options would be about 9 percent While this is a somewhat hazardous calculation because of unspecified ability differences, the calculation does suggest that a training-options interpretation is at least plausible

---

[7] The distribution is skewed partly because for a few persons the implied negative years of predegree experience were set to zero.

Utilizing the earnings function given above, we rely on experience-earnings profiles as estimated from cross-sectional data to investigate the role of graduate school quality on lifetime earnings, to compare economics with other disciplines over time, and to develop other topics Given this use of cross-section data, it is well to keep in mind some of the limitations of such approximations to the lifetime profiles Not only will the profiles shift up over time (as we will discuss in Section V), but there may be considerable heterogeneity of individuals distributed by age (and hence by experience) as a result of factors such as promotion policy [9] To

---

[8] The rate of return to choosing the training options associated with the typical individual with *XPR* = 5 was calculated by finding the discount rate (r), which satisfies

$$\sum_{XPO=0}^{34} [S\hat{A}L(XPO,\ XPR = 4) - S\hat{A}L(XPO,\ XPR = 6)](1 + r)^{-XPO} = 0$$

[9] Younger economists as of 1970 (say, with less than 10 years experience on the job) had acquired a knowledge of theory and econometrics which, by and large, economists with, say, 20 years of postdegree experience did not acquire If such skills are earning quasi rents as of 1970, our profiles will generally be too flat relative to a long-run equilibrium profile

TABLE 2

PREDICTED SALARIES OF PH.D. ECONOMISTS BY SCHOOL OF DEGREE AND YEARS OF POSTDEGREE EXPERIENCE

SCHOOL	YEARS OF POSTDEGREE EXPERIENCE*							
	0	5	10	15	20	25	30	35
Chicago	12,610	15,490	18,290	20,770	22,690	23,830	24,070	23,370
Columbia	11,230	14,280	17,230	19,720	21,410	22,070	21,580	20,020
Cornell	11,826	14,060	16,200	18,090	19,580	20,520	20,850	20,520
Harvard	11,690	14,670	17,500	19,850	21,390	21,920	21,350	19,760
Yale	10,720	14,590	18,340	21,300	22,860	22,680	20,780	17,600

* Years of predegree experience assumed to be 5.

emphasize these kinds of difficulties, we have estimated, for graduates of several individual schools, regressions of earnings on our experience and other variables in the form of equation (2) and have presented predicted salaries from these regressions (table 2). While the estimated profiles for most of the schools look reasonable, the Yale profile is curious, peaking out at 20 years of postdegree experience. In part this reflects the fact that Yale graduates have had very few years of predegree experience (2.3 years), and to fit the early (steep) part of the profile the quadratic form imparts some distortion to the end of the experience-earnings profile.

III. Effect of Quality of Graduate School on Salaries

Although it is customary to assume that labor is homogeneous, it is commonly recognized that members of particular occupations vary widely with respect to ability, and this is true of professional labor markets as well. In this section we investigate how one proxy for ability, the "quality" of the graduate school attended, influences lifetime earnings of academic economists. It would be desirable, of course, to have direct measures of each individual's productivity (e.g., number and quality of publications, teaching effectiveness, and the like), but these are difficult to obtain for a large sample survey and are not available in the NSF data set. The NSF data do include the name of the graduate school the respondent attended, so we were able to match this information with the periodic surveys of the effectiveness of graduate programs. We then constructed a one-zero dummy variable denoting receipt of doctorate from a highly ranked graduate school, in particular the top 10 (*RNKD*).[10] The influence of this variable can be attributed to two potentially separate factors, the effectiveness of the graduate program itself and the selectivity of the program with respect to admissions Although this raises a very interesting question in itself, there is no conceivable way we could identify the individual influence of each factor with the present data set, and we thus employ *RNKD* primarily to reduce the heterogeneity of the sample

---

[10] Our procedure was as follows. Those respondents who received their doctorates prior to 1940 were assigned the ranking from the Hughes study of 1925, for graduates in the 1940–59 period, the rankings of the 1957 Keniston study, and for 1960–70 graduates, the 1964 Cartter ranking. In this way the individual's rank depends on both the school attended and his vintage. Those schools with top-10 ranks were combined into a single dummy variable, *RNKD*. The figure 10 was chosen arbitrarily—determined in part by the availability of data in other academic professions. These data are found in Cartter (1966).

Introduction of the quality of graduate school variable yielded the following results

$$\begin{array}{lllll} \ln SAL = & 0431XPO & -\ 000750(XPO)^2 & +\ 0111XPR & -\ 000203(XPR)^2 \\ & (\ 0013) & (\ 000036) & (\ 0021) & (\ 000090) \\ & -\ 000679XPO \times XPR & -\ 136FEM & +\ 057RNKD & +\ 4\ 717 \\ & & & & (\ 009) \end{array}$$

$$R^2 = 468, \text{ standard error of estimate } = 192 \qquad (3)$$

This indicates that receipt of a doctorate from a graduate school ranked in the top 10 increases an economist's 9-month salary by an average of about 6 percent [11] Of greater interest are the results in which *RNKD* is interacted with *XPO* and $(XPO)^2$ These are

$$\begin{array}{lllll} \ln SAL = & 0362XPO & -\ 000585(XPO)^2 & +\ 0095XPR & -\ 000170(XPR)^2 \\ & (\ 0015) & (\ 000043) & (\ 0021) & (\ 000089) \\ & -\ 000578XPO \times XPR & -\ 135FEM & -\ 059RNKD & \qquad (4) \\ & (\ 000099) & (\ 016) & (\ 015) & \\ & +\ 0200RNKD \times XPO & -\ 000508RNKD \times (XPO)^2 & +\ 4\ 756 & \\ & (\ 0024) & (\ 000074) & (\ 013) & \end{array}$$

$$R^2 = 481, \text{ standard error of estimate } = 189$$

These results suggest that there are sharp differences in the shapes of the salary/experience profiles between economists who received their graduate training from the top 10 graduate schools and those economists who attended less prestigious graduate schools The predicted starting salaries of the latter are over 5 percent greater than the starting salaries of the former, but the rate of growth of salaries is much greater in the early stages of the work cycle for graduates of the top 10 schools than for others [12] The same pattern has been found in physics, anthropology, sociology, and mathematics In biology this pattern was not upheld, and alternate measures of school quality (e g , 1–0, whether graduated from Berkeley or Harvard) did not yield any indication that graduate school attended influenced lifetime earnings The results for economists are

---

[11] The percentage wage advantage accruing to graduates of ranked schools is 100 × $(e^b - 1)$, where b is the coefficient on *RNKD* in the regression explaining log *SAL*

[12] The *F*-statistic on the two interaction terms is 37 7, which is highly significant

TABLE 3

PREDICTED SALARIES OF ACADEMIC ECONOMISTS FOR PH.D.'S BY RANK OF GRADUATE SCHOOL AND BY YEARS OF POSTDEGREE EXPERIENCE (IN DOLLARS)

	YEARS OF POSTDEGREE EXPERIENCE*							
SCHOOL RANK	0	5	10	15	20	25	30	35
Top-10 rank	11,460	14,550	17,490	19,910	21,460	21,900	21,160	19,360
Other	12,150	14,140	15,980	17,540	18,700	19,350	19,460	19,000
Top/other†	0.943	1.029	1.094	1.135	1.148	1.131	1.087	1.019
	(.015)	(.010)	(.011)	(.013)	(.014)	(.015)	(.017)	(.022)

* Years of predegree experience assumed to be 5.
† Estimated standard error of ratio in parentheses.

portrayed in table 3 in which predicted salaries for *XPR* = 5 (i.e., starting age $\cong$ 30) are shown. The salary ratio of an individual with 25 years of postdegree experience (in this case about age 55) to a new entrant is 1.91 for graduates of schools ranked in the top 10 versus 1.59 for graduates of other schools. While the *RNKD* variable does interact with postdegree experience terms, it does not interact substantially with the predegree experience terms (*F* = 2.4). This is not surprising, for the payoff to a more elaborate graduate program is likely to accrue to what one does after completion of the degree.

From our comparison of postdegree experience profiles of Ph.D. graduates from ranked and other programs, we infer that the differential shapes reflect the fact that more highly trained and/or able persons engage in on-the-job training more extensively than less able or less highly trained individuals. Since this inference is wholly dependent on the shape of the earnings profiles, we ran separate regressions for individuals who attended schools ranked 1–5, 6–10, 11–16, and all others as an additional piece of evidence on the relation between a partial measure of skill and the life-cycle path of earnings. The results indicate that the salary/experience profiles get progressively less steep from the 1–5 group to the unranked schools. We will not report the regression results (which are based on the form of eq. [2]), but for individuals who began their professional experience at about age 30 (*XPR* $\cong$ 5), the estimated ratio of a person with 20 years of postdegree experience to a new entrant is 1.91 for the 1–5-ranked schools, 1.79 for 6–10, 1.70 for 11–16, and 1.57 for those ranked 17 or lower. At the same time, the average academic salary differentials between graduates of the most and least prestigious institutions are not very large. As table 3 shows, Ph.D. graduates from the top 10 schools start out at a salary (6 percent) less than graduates from schools other than the top 10.[13] Salaries are roughly equal at about 3 years of professional experience, and the differential slowly increases to a 15 percent advantage after 20 years of experience. The average percentage salary advantage accruing to graduates from the top 10 universities is about 8 percent over a working life. Because the largest relative benefits come relatively late, however, the relative present values at the time of entry to the work force are closer than is implied by this 8 percent figure.

---

[13] A similar result has been noted by Francis M. Boddy in his annual surveys of the larger (Ph.D.-granting) departments (Boddy 1971). A widening experience differential across ability groups is reported in Hause (1971).

One interesting comparison is between the results we have portrayed and those reported in the Tolles and Melichar volume (Tolles and Melichar 1968). Using 1966 median salary and reported years of experience, the estimated ratio of peak to beginning salary is about 1.67 for Ph.D.'s in academic employment. While this generally compares to the average for our 1970 data, according to our profiles for $YPR = 5$, the disaggregation of academically employed into different values of $RNKD$ and XPR provides a measure of the extent to which profiles vary in terms of shape. At one extreme there are early starters ($XPR = 0$) from schools ranked 1–5 who have a ratio of peak to entry salary of 2.14, and at the other extreme late starters from schools ranked ≥ 17 who have a ratio of peak to entry salary of 1.48. That is, the various parts of the academic labor market for Ph.D.'s vary widely with respect to the extent to which there is a strong dependency between years of experience and salary as well as large differentials by degree level.[14]

IV. Determinants of Academic Salaries in Different Types of Institutions

To a certain extent the differential effects of $RNKD$ on salaries may be attributed to differences in type of employment between economists from ranked and unranked graduate schools. It is also well known that salaries at such institutions are relatively low for lower ranks and relatively high for higher ranks. Hence, adjusting for the type of institution at which the individual is currently employed may explain a large part of the relation between salaries and graduate school rank. A human-capital interpretation of academic earnings would predict that there would be systematic differences in the earnings structure between institutions which specialize primarily in teaching versus those which specialize in research. The argument is that research and thesis-supervision skills generally require a more extended period to develop than teaching skills, so that schools with a research orientation will have earnings profiles which fan

---

[14] The differential between Ph.D. and master's degree holders is 11 percent at 0 years of experience and rises to 51 percent at 25 years of experience for graduates of top-ranked schools. The corresponding figures are 35 and 52 percent, respectively, for graduates of other programs. The rise in this differential by years of experience provides additional evidence on the life cycle of acquired skill and earnings wherein more "able" individuals train for longer periods of time and thus realize a greater experience (and age) dependency to their salaries.

TABLE 4

Determinants of 9-Month Salaries for Academic Economists for Selected Groups of Schools, 1970*

Variable	(5) Entire Sample	(6) Top Twenty	(7) Small Private	(8) Deep South	(9) Great Plains	(10) University of California	(11) California State
XPO	043	055	041	020	038	049	053
	(001)	(003)	(001)	(005)	(007)	(005)	(006)
$(XPO)^2$	– 00075	– 00089	– 0007	– 00047	– 00074	– 00072	– 0012
	(00004)	(00007)	(0003)	(00016)	(00021)	(00017)	(00017)
XPR	011	013	– 007	– 001	002	022	007
	(002)	(007)	(023)	(007)	(011)	(014)	(014)
$(XPR)^2$	– 00020	– 00003	– 00154	– 00016	– 00044	– 00066	– 00090
	(00009)	(00040)	(00129)	(00022)	(00062)	(00088)	(00078)
XPO × *XPR*	– 00068	– 00056	– 00086	– 00006	– 00011	– 00032	00044
	(00010)	(00032)	(00107)	(00034)	(00069)	(00056)	(00060)
FEM	– 14	– 18	– 09	– 16		– 44	– 10
	(02)	(04)	(17)	(08)		(11)	(06)
RNKD	06	– 01	04	– 07	– 01	01	– 02
	(01)	(02)	(06)	(06)	(07)	(03)	(04)
Constant	4 72	4 00	4 72	4 90	4 77	4 66	4 68
	(0 01)	(0 02)	(0 09)	(0 04)	(0 05)	(0 03)	(0 06)
R^2/S E E	47/ 19	69/ 18	37/ 23	24/ 14	49/ 15	79/ 14	72/ 14
Sample size	3,040	411	66	123	74	109	64

* Given the small number of females in the samples the corresponding coefficients should be interpreted with caution

out in comparison with schools which emphasize teaching. To see if this is the case, we created six subsamples of academic economists (1) those *Ph.D. economists who are employed by the schools in the top 20 of the* 1964 Cartter rank, (2) a selection of 34 small private colleges of reputedly high quality,[15] (3) schools in the Deep South (Alabama, Arkansas, Louisiana, Mississippi), (4) schools on the Great Plains (the Dakotas, Wyoming, Nebraska, Kansas, Iowa, and Oklahoma), (5) all schools in the University of California system, and (6) all schools in the California State College system.

The regression results for each of these subsamples (as well as eq. [2] renumbered as [5] for comparative purposes) are presented in table 4. We have also reported the means of the variables by school type in table 5 and predicted salaries by years of postdegree experience in table 6. There are several notable results.

First, there is a wide range in the values of the estimated parameters on the postdegree experience variables. The estimated coefficient on XPO, *which is the rate of growth of salaries in cross-sectional terms at* $XPO = 0$, ranges from .020 for the Deep South subsample to .055 for the top-20 subsample. Accordingly, the estimated ratio of individuals with $XPO = 20$ to those with $XPO = 0$ varies from 1.23 for the Deep South to 2.00 for the top 20. Substantially the same pattern across school types was obtained in the other disciplines for which we developed such comparisons (biology, mathematics, and physics) and over time (1964 and 1970). To some extent this difference in profile steepness across schools reflects selective-retention policies with respect to junior faculty and is another example of cohort heterogeneity mentioned in Section II. Yet (as developed below) we know that differences in profile steepness relate to ranking of graduate school, and graduate school attended influences type of school of employment. As argued in our interpretation of graduate school quality, the profile shapes represent differences in the life cycle of learning *on the job and renting out one's accumulated skills*.

Second, the estimated coefficient on $RNKD$ is not significantly different from zero in each of the six subsamples; indeed, the point estimates are negative in four of the six. However, in the school groups with a large fraction of individuals who hold Ph.D.'s from the top graduate schools, the estimated experience-earnings profiles are much steeper than in the schools which generally hire from other than the top 10 graduate schools. These results suggest that the principal reason for the positive value of $\partial^2 (\log SAL)/\partial(XPO)\,\partial(RNKD)$ observed in equation (4) is that

---

[15] Amherst, Antioch, Bates, Beloit, Bowdoin, Bryn Mawr, Carleton, Colgate, Connecticut College, Davidson, Dennison, Furman, Goucher, Hamilton, Haverford, Harpur, Kalamazoo, Middlebury, Mills, Mt. Holyoke, Oberlin, Occidental, Reed, Rollins, Sarah Lawrence, Smith, Stephens, Swarthmore, Trinity, Wellesley, Wesleyan, Williams, Wooster, and Vassar.

TABLE 5

MEANS OF VARIABLES IN 1970 SUBSAMPLES

Variable	Entire Sample	Top Twenty	Small Private	Deep South	Great Plains	University of California	California State
SAL	15,890	18,560	15,980	14,750	14,830	16,270	15,300
*INC**	20,570	24,590	19,070	19,020	19,950	22,170	19,020
RFS†	0.66	0.82	0.56	0.51	0.54	0.92	0.55
XPO	9.7	11.0	11.1	9.2	8.8	9.7	9.7
XPR	4.3	2.5	3.9	6.4	4.3	2.2	5.3
FEM	.05	.03	.03	.02	0.00	.02	.13
RNKD	.33	.65	.50	.05	.08	.70	.25
Sample size	3,040	511	66	123	74	109	64

* Total annual professional income

† Proportion reporting research as a major work activity

TABLE 6

PREDICTED 9-MONTH ACADEMIC SALARIES OF ECONOMISTS FOR SELECTED GROUPS OF SCHOOLS, 1970

XPO	Entire Sample	Top Twenty	Small Private	Deep South	Great Plains	University of California	California State
0	11,770	11,680	11,230	13,310	11,800	11,520	10,870
5	14,090	14,850	13,260	14,520	13,950	14,340	13,900
10	16,250	18,060	15,070	15,460	15,870	17,200	16,750
15	18,050	21,010	16,490	16,080	17,410	19,910	19,050
20	19,310	23,370	17,380	16,330	18,400	22,230	20,430
25	19,890	24,870	17,640	16,190	18,730	23,940	20,660
30	19,750	25,320	17,230	15,680	18,370	24,860	19,710
35	18,880	24,650	16,200	14,840	17,360	23,910	17,730

NOTE.—*XPR* = 5 and *RNKD* = 0.

graduates of the more prestigious schools are more likely to receive appointments in universities with a research orientation rather than at schools which emphasize teaching.

Third, the results for the school types are also useful in interpreting the breakdown of total experience into its pre- and postdegree components. The estimated coefficients on *XPR* and XPR^2 in the six subsamples, like those on *RNKD*, vary rather widely, but on average the estimated effect on *SAL* in the subsamples is smaller than for the entire sample. Notice that across the subsamples the mean value of *XPR* is inversely related to the steepness of the estimated age-earnings profile. One possible explanation of these results is that individuals who have a large amount of predegree experience are often those who had difficulties finishing their doctoral dissertations and took jobs at schools which specialize in teaching prior to completion of their degrees. Although they eventually finish their degrees, they still specialize in teaching (because they probably were not strongly attracted to research in the first place) so that their research abilities—and hence salaries—did not grow very quickly If this interpretation is correct, then the *XPR* variable, like the *RNKD* variable, reflects factors by which individuals are sorted into employment at different types of institutions [16]

Fourth, the average values of *RNKD* and *XPR* for the small-private subsample are fairly close to those for the top-20 sample Using *RNKD* and *XPR* as proxies for ability, the differences in the level and shape of earnings profiles between the two can be taken as indicative of lifetime earnings differences between those who choose to emphasize research versus those who emphasize teaching The estimated salary advantage of individuals in the top 20 over those from the small-private group starts out at 4 percent at *XPO* = 0 and rises monotonically to 52 percent at the time of retirement There are, of course, a number of difficulties with interpreting this differential as solely attributable to differences in research behavior (e g , nonpecuniary advantages and disadvantages at both sets of schools at different stages of the work cycle), but we expect that the differential would behave qualitatively in this fashion

In Sections II–IV we have presented the basic features of lifetime earnings patterns of academic economists Qualitatively, the results obtained for economists were found to hold with virtually no exceptions in the other disciplines which we have studied, and some comparisons have been given We now turn to comparison of the structure of earnings in economics to that of other disciplines and over time

V. Earnings in Economics and Other Disciplines, 1964 70

The structure of economists' and sociologists' earnings with respect to postdegree experience was subject to rather minor change from 1964 to 1970, but in biology, mathematics, and particularly in physics there were substantial changes in the structure of earnings To make these interfield comparisons, we estimated equation (2) for five disciplines in both the 1964 and 1970 NSF Register data The estimated experience-earnings function of economists shifted up over this 6-year interval by a constant

---

[16] These contentions are supported by the following regression equation

$$\begin{aligned} TOP20 = {} & \underset{(017)}{240}\,TOP5 + \underset{(024)}{155}\,TOP(6\text{–}10) + \underset{(027)}{114}\,TOP(11\text{–}16) \\ & - \underset{(003)}{024}\,XPR + \underset{(0002)}{0007}\,(XPR)^2 + \underset{(013)}{188}, \\ & R^2 = 10, \text{ standard error of estimate } = 38 \end{aligned}$$

Here *TOP*20 = whether employed in top 20, *TOP*5 = whether attended a school ranked 1–5, *TOP*(6–10) = whether attended a school ranked 6–10, and *TOP*(11–16) = whether attended a school ranked 11–16 The excluded group is those graduating from all other schools proportionate amount (about 50 percent) for all relevant values of XPO In 1964 the predicted 9-month salary of the "typical" economist (XPO = 12 and XPR = 4) was $11,320, in 1970 the predicted salary of this same individual (XPO = 18 and XPR = 4) was $18,770, an increase of 66 percent over the 6-year period[17] This 66 percent increase can be attributed to three sources (1) price inflation, which was 25 percent during this period, (2) real wage growth of economists, which was about 20 percent, and (3) a movement up (or down) the cross-sectional experience-earnings profile Sources (1) and (2) account for the 50 percent upward shift in the profile from 1964 to 1970 (1 25 × 1 20 = 1 50), and this typical individual would be expected to gain 11 percent in salary due to moving from a postdegree experience level of 12 to 18 Thus, for the typical academic economist in the beginning period, approximately 0 35 of his money wage gains from 1964 to 1970 were due to real wage growth for the profession as a whole, 0 45 were due to price inflation, and 0 20 were due to an improved position in the life cycle

These results for economists and the other disciplines are presented in table 7 The percentage rates of salary increase for "typical" members of each discipline (XPO = 12 in 1964 and XPO = 18 in 1970) were as follows sociology, 56, biology, 63, physics, 63, and mathematics, 66 However, the relative gains of young and old cohorts differ rather dramatically across the disciplines When we examine table 7C, the percentage change of predicted salaries of new entrants (XPO = 0) varies from 50 percent in economics to 26 percent in physics Upon deflating these by the rise in consumer prices, we see that the starting real salary of Ph D economists rose by 20 percent but remained virtually unchanged for Ph D physicists This differential growth is also reflected in cohort comparisons New entrants (XPO = 0) in economics and sociology realized a rise in nominal salary of 87 percent and 84 percent, respectively, after 6 years of experience, but new entrants in physics and mathematics realized a 72 percent and 80 percent rise in nominal salary, respectively In contrast, senior scholars (XPO = 24 in 1964) in economics and sociology realized a rise in nominal salary of 48 and 46 percent, respectively, after 6 more years of experience, whereas senior scholars in physics and mathematics realized a 48 and 51 percent rise in nominal salary, respectively

What factors explain these interfield differences in changes in the structure of compensation? Although we have a very limited sample (the data points are very expensive), we can offer some comments on supply-and-demand influences on the compensation of different experience cohorts in the fields In table 8 we have summarized information on

---

[17] An XPR of 4 was used because it is closer to the mean XPR across the six fields studied

TABLE 7

ACADEMIC SALARIES FOR SELECTED DISCIPLINES BY YEARS OF POSTDEGREE EXPERIENCE, 1965 AND 1970

A PREDICTED SALARIES, 1970

Postdegree Experience	Economics	Sociology	Biology	Physics	Mathematics	Anthropology
0	11,560	10,650	9,790	9,470	10,560	10,460
6	14,340	13,370	12,420	12,940	13,850	13,260
12	16,850	15,760	14,850	16,330	16,940	15,980
18	18,770	17,460	16,740	19,030	19,330	18,290
24	19,800	18,150	17,790	20,470	20,560	19,900
30	19,790	17,730	17,830	20,330	20,400	20,508
36	18,750	16,260	16,850	18,640	18,870	20,210

B PREDICTED SALARIES 1964

Postdegree Experience	Economics	Sociology	Biology	Physics	Mathematics
0	7,720	7,280	6,980	7,540	7,700
6	9,610	9,010	8,700	9,710	9,800
12	11,320	10,550	10,260	11,660	11,660
18	12,630	11,690	11,460	13,080	12,970
24	13,340	12,250	12,100	13,700	13,480
30	13,350	12,150	12,110	13,390	13,090
36	12,640	11,410	11,420	12,200	11,880

C PERCENTAGE SHIFT IN ESTIMATED PROFILES, 1964–70

Postdegree Experience	Economics	Sociology	Biology	Physics	Mathematics
0	49.7	46.3	40.3	25.6	37.1
6	49.2	48.4	42.8	33.3	41.3
12	48.9	49.4	44.7	40.1	45.2
24	48.4	48.2	47.0	49.4	52.5
30	48.2	45.9	47.2	51.8	55.8
36	48.3	42.5	57.5	52.8	58.8

senior/junior salary differentials (relative wages), the proportion of Ph.D.'s with 14 or fewer years of experience (relative supplies), and the rate of growth of undergraduate students enrolled (shifts in demand). If one interprets junior and senior faculty as the two inputs used in producing instructional services and assumes a homothetic production function, then by regressing relative supplies on relative wages an estimated substitution elasticity of 5.6 is obtained.[18] This suggests that junior and senior faculty

---

[18] The estimated regression equation was

$$\log \frac{L(XPO \geq 14)}{L(XPO < 14)} = \underset{(.28)}{2.02} - \underset{(.76)}{5.56} \log \frac{SAL(XPO \geq 14)}{SAL(XPO < 14)},$$

where $R^2 = .95$

TABLE 8

SENIOR/JUNIOR SALARY DIFFERENTIALS, RELATIVE SUPPLIES, AND ENROLLMENT GROWTH

Field	$\frac{\text{Salary } (XPO = 24)^*}{\text{Salary } (XPO = 6)}$	$\frac{L(XPO \geq 14)^\dagger}{L(XPO < 14)}$	% Change in B.A.'s, 1964–70
Economics	1.38	1.27	66
Sociology	1.36	1.44	177
Biology	1.43	1.04	78
Mathematics	1.48	0.75	48
Physics	1.58	0.64	6

* Salary estimates from table 7A.

† National Science Foundation. *American Science Manpower* (1968), table A-5.

are quite substitutable in the academic production process. Another possibility is that differences in relative salaries by age are caused by different rates of growth of demand for different disciplines. When the growth variable is added to the regression explaining relative demands, its estimated coefficient is positive but statistically insignificant ($t = .48$), and the estimated value of the elasticity of substitution between older and younger faculty falls only to 4.9. Yet there is certainly a strong simple correlation between undergraduate demand and relative salaries ($r = -.83$).

Another factor to be considered in the differential changes in the salary/experience profiles is the differential acceleration in Ph.D. output in the fields, which is measured by the ratio of senior to junior scholars. By this measure physics and mathematics enjoyed the biggest boom in the 1960s, and this would encourage rapid salary growth since established scholars train Ph.D. and master's students. As a result, the salary structure in 1970 will reflect (1) the large gains realized during the decade by those in the discipline prior to the early 1960s, coupled with (2) the slack in demand in the undergraduate sector which lowers the compensation for new positions. In addition, a large increase in the supply of new Ph.D.'s should bid up the implicit price for training options (particularly in the short run) and lower salaries for those in the early career stages. To summarize, these factors—namely, changes in relative supplies, differential growth in demand for graduate training during the 1960s, and differential growth in undergraduate enrollment—have operated to effect differential changes in the salary structure across the fields. Given our small sample of fields, the relative importance of the different factors cannot be isolated.

In our discussion of earnings across fields we have focused solely on the academic sector. This allows us to proceed without measures for a number of nonpecuniary factors which are difficult to quantify and also allows us to compare standardized 9-month salaries. If comparisons are made

TABLE 9

Gross Professional Income by Sector of Employment and Discipline, 1970

A Annual Earnings in Academics, 1970

XPO	*ECON*	*SOC*	*MATH*	*BIO*	*ANTHR*	*PHYS*
5	17,550	14,730	16,370	14,700	15,360	15,130
15	23,340	19,420	22,770	19,780	20,770	22,250
25	25,900	20,770	25,530	22,260	24,470	25,540

B Annual Earnings in Government, 1970*

XPO	*ECON*	*SOC*	*MATH*	*BIO*	*PHYS*
5	19,320	17,990	18,850	14,890	17,850
15	24,560	22,240	24,900	20,680	24,680
25	26,390	23,120	27,650	24,010	27,860

C Annual Earnings in Business, 1970†

XPO	*ECON*	*MATH*	*BIO*	*PHYS*
5	24,810	21,910	18,110	23,420
15	33,770	27,930	24,140	26,310
25	37,110	29,650	27,730	29,430

* Insufficient sample in *ANTHR*
† Insufficient sample sizes in *SOC* and *ANTHR*

across the disciplines in other employment sectors (i e , government and business), we would expect that the relative salaries of the different disciplines would be quite similar to academics because of within-field migration which would tend to equalize net advantages In table 9 we present a comparison of gross professional income (12-month earnings) in academics, government, and business in 1970

Annual earnings in both academic and government employment (at *XPO* = 15, for example) are rather comparable in economics, physics, and mathematics, but are lower in biology and sociology

In business employment, Ph D economists do markedly better than the other disciplines However, business earnings are generally highest if one develops managerial skills, and these are generally learned at the expense of more narrowly defined professional abilities in the disciplines The evidence which we have on this point is that for economics and physics, the fields with the greatest business employment, there is a reverse temporal pattern to the earnings advantage of graduating from a highly ranked school relative to the pattern observed in academics initially the earnings advantage of graduates of highly ranked programs is substantial (economics, 14 percent, physics, 6 percent), but the earnings advantage converges to zero at high levels of postdegree experience

VI. Summary and Conclusion

In this paper we have employed the NSF National Register data to examine the earnings of academic economists from a number of points of view Among our major conclusions are the following

1 An individual who completes his doctorate at a relatively early point in the life cycle has a markedly different earnings pattern from an individual who begins his career at a later point in the life cycle Early starters take positions with lower starting salaries but which afford greater human-capital accumulation possibilities than the positions typically taken by late starters Consequently, the ratio of peak to beginning salaries is much greater for early starters than for late starters (1 88 for economists who received the Ph D at about age 25 compared to 1 58 for economists who received the Ph D at about age 35) The estimated rate of return to on-the-job training for economists was found to be 7 percent

2 A second way of differentiating the academic labor force is by the quality of the graduate school attended We found that those economists who earned the doctorate from one of the top 10 graduate schools receive lower initial salaries than economists from less prestigious schools, but the rate of salary increase tends to be much greater for the former than the latter The predicted ratio of peak to beginning salaries is 1 91 for graduates of the top schools and 1 60 for other economists It would be interesting to know to what extent these differences in life cycle-human capital accumulation patterns can be attributed to the content of graduate schools versus the greater selectivity of the top graduate schools, but this question could not be answered with the NSF data set

3 We also analyzed the determinants of the salaries of economists employed in selected types of schools—the top 20 Cartter-ranked schools, a set of selective small private colleges, all institutions in the Deep South, all institutions on the Great Plains, the California State College system, and the University of California system As would be expected, the more research-oriented school types exhibit much steeper experience-earnings profiles than the less research-oriented school types It is also true that the research mills have a higher proportion of their faculties who are early starters and who received the doctorate from a top university (They are also less likely to employ women than are institutions which stress teaching) These facts suggest the institutional reasons for our results with respect to early versus late starters and economists from the better versus other graduate schools

4 The salaries of academic economists in 1970 were not much different from those of physicists, mathematicians, and anthropologists and only a little higher than those of sociologists One exception to this is the fact that *economists earn relatively higher starting salaries than anyone else in our selection* However, gross earnings were greater for economists than for

any of the other disciplines studied These same findings also applied in 1964

5 The analysis of earnings in various disciplines also provides a suggestion of a finite elasticity of substitution between labor of different ages Specifically, disciplines in which a large fraction of the occupation has high amounts of experience are likely to be characterized by relatively flat experience-earnings profiles as compared to disciplines in which a small fraction of the occupation has high amounts of experience The estimated value of the elasticity of substitution between labor with 14 or more years of experience and labor with fewer than 14 years of experience is about five This implies that to some extent the demand for educated labor is age-specific, and models of the dynamics of professional labor markets (Freeman 1971) should take this into account

References

American Association of University Professors "Coping with Adversity Report on the Economic Status of the Profession " *American Assoc Univ Professors Bull* 58 (Summer 1972) 178 243

Ben-Porath, Yoram "The Production of Human Capital and the Life Cycle of Earnings " *J P E* 75 (July/August 1967) 352-65

Boddy, Francis M "Recent Behavior of Economists' Salaries ' *A E R* 61 (May 1971) 316 20

Cartter, A M *An Assessment of Quality in Graduate Education* Washington American Council on Educ , 1966

Freeman, Richard B *The Market for College Trained Manpower* Cambridge, Mass Harvard Univ Press, 1971

Hause, John "Ability and Schoolings as Determinants of Lifetime Earnings or If You're So Smart, Why Aren't You Rich?" *A E R* 61 (May 1971) 289-98

Johnson, George E , and Stafford, Frank P 'The Earnings and Promotion of Women Faculty " *A E R* (1974), in press

Mincer, Jacob "On-the-Job Training Cost, Returns, and Some Implications " *J P E* 80 (suppl , October 1962) 50–79

—— - *Schooling, Experience, and Earnings* New York Nat Bur Econ Res , 1973

Mooney, Joseph D "Attrition among Ph D Candidates An Analysis of a Cohort of Recent Woodrow Wilson Fellows " *J Human Resources* 3 (Winter 1968) 47-62

Rosen, Sherwin "Learning and Experience in the Labor Market " *J Human Resources* 7 (Summer 1972) 326–42

Stephan, Paula E "Life Cycle Training Paths Subsequent to Formal Education " Ph D dissertation, Dept Econ , Univ Michigan, 1971

Stigler, George J *Capital and Rates of Return in Manufacturing* New York Nat Bur Econ Res , 1964

Tolles, N Arnold, and Melichar, Emanual "Studies of the Structure of Economists' Salaries and Incomes " *A E R* , vol 58 (December 1968, pt 2)

Warkov, Seymour, et al *Graduate Student Finance, 1963* Chicago Nat Opinion Res Center, 1965

Frank Knight in Retrospect

Frank Knight and the Tradition of Liberalism

Scott Gordon

Indiana University and Queen's University

Frank Knight was a very complex and subtle thinker, especially so in the area of political philosophy, and any appreciation of his views in short compass (or perhaps even at length) is bound to produce simplifications of the kind that would have made him groan In the 60 years of his study and meditation upon man as a political animal (i e , a rational one, not a behavioristically social one like the termites or ants), he generated far more questions than answers, more problems than solutions, and he was more distressed by ingenuousness in political philosophy than by error or evil He had the uncommon gift (and the curse!) of the compound eye, human society appeared to his perception through many angles of view which compelled a projection that was a mosaic of great richness, complexity, and, ultimately, mystery We may label him a "liberal" (in the midnineteenth-century meaning of that term) because he did so describe himself, but his own discussion of liberalism was more concerned with its weaknesses than its strengths No one, indeed, has searched out its subterranean defects with more tenacity

The essence of the liberal philosophy, in Knight's view, was freedom, not as an instrument (for utilitarian efficiency) or even as a human preference, but as an ethical value in itself This raises many philosophical problems, which are compounded by the fact that Knight was unwilling to accept any specific theory of ethics or even any of the extant general approaches to ethical theory He rejected alike all reductionist theories (including the reduction of all values to freedom) and all deontological statements of absolute values The closest he would come to advancing an ethical position of his own was to evince a belief (or a hope) that man is capable of attaining higher levels of social existence—to become more "civilized" in some sense Yet he rejected altogether the idea that one may describe some ideal civilization and set it up as a goal, the ceaseless search for the better being superior to any conceivable "best " It would ill serve the purpose of even a brief essay on Knight's political philosophy to focus exclusively upon his liberalism, since the problem of freedom was intimately connected in his thought with a number of other issues that are equally fundamental We will do better to circle around our target, as Knight himself always did, in order to note the vital intersections between freedom and other facts and values

We may begin where Knight did, with uncertainty His doctoral dissertation, *Risk, Uncertainty, and Profit* (written in 1914–16, published in 1921) can be (and has been) read as a technical study on the pure theory of profit as a category of income distribution More broadly, it can be read as an essay on the nature of a private enterprise economy But it is more than these If one does not neglect those passages which may appear to the economist as *obiter dicta,* it becomes clear that Knight was tackling one of the most fundamental problems of epistemology and ethics In Knight's view, uncertainty is not only, or "merely," the key to the understanding of an economy of private entrepreneurship whose motive force is profit, it is the most basic characteristic of human existence, which has profound philosophical and practical significance and produces much of the array of political and ethical problems which man faces as a participant in social organization A world lacking in uncertainty (even one in which "risk" were present) would not be one of free human beings at all, but of determined automata In such a world, the steady forces of biological evolution would inexorably fix all survival-value behavior patterns in the genetic inheritance No meaning could then attach to words such as "deciding" or "valuing", indeed, consciousness itself would be redundant, and "thinking" would not be distinguishable from "behaving" or reacting to stimuli A Laplacean world leaves nothing to be decided by acts of will Only a world of uncertainty can be modified by the values man adopts and the choices he makes, and it is man's capacity to change the world, and himself, that is the most essential quality, and opportunity, of the human condition Yet, at the same time, man's opportunity of acting to advance his civilization must be based upon reason (his capacity to examine the world objectively and to establish cause and effect) which is inherently limited by uncertainty itself Thus, man is a free and thinking being because of uncertainty, yet it is uncertainty which imposes limits upon his effective use of reason, a complexity that is compounded by the fact that we are uncertain also as to the limits of uncertainty, there is no way to establish the boundaries among knowledge, opinion, and plain ignorance

The most important implication of this, for liberalism and for economics, is that man is a fluid creature and the course of his psychological evolution is indeterminate In Knight's view, the largest defect in liberalism as a social philosophy is that it takes the individual as a datum, and defines the objective of social policy as ministering to his given wants, this turns a blind eye to the phenomenon of acculturation, the fact that the individual and his wants are largely formed by the social process and are not exogenous to it One cannot set up as an ethic that it is good that people should have as much as possible of what they want, because the ethical quality of a civilization is largely a matter of what its people want and the direction in which those wants are evolving under the impact of choices successively made Knight was the only economist of the twentieth century to take up seriously the problem raised by Alfred Marshall of the complex relation between man's "wants" and the "activities" he undertakes to satisfy them The growth of economic wealth, in removing man from the pressure of immediate physical necessity creates a society in which wants are endogenous to the social process, neither ends nor means can be taken as given, nor are they separable from one another The evolutionary character of human culture as a consequence of this is what Marshall was referring to in his rather cryptic remarks about the necessity of economics becoming a "biological" science Like Marshall before him, Knight could not provide an analytical framework that could grapple with this problem but, unlike Marshall, he was not naïve enough to believe that it will ever yield to the advancement of "science" in any meaningful sense of that term As a philosopher, Knight was unwilling to accept any conception of an ideal as a guide to human evolution, or any teleological belief that society tends toward some goal On this issue, as on many others, he arrived at doubt rather than doctrine, an increased confidence only in uncertainty

A connected theme, which appears time and again in Knight's essays, is the significance of play in human culture This is a phenomenon of extraordinary significance for two reasons First, the means-ends relationship is reversed in play activity, goals are established as instrumental to the true objective, which is to provide opportunity for satisfying activity Second, a good game requires rules, yet there is an inherent conflict between the motivation to win and obedience to the rules What Knight wished to point out is that many economic (as well as political and other social) activities have the essential character of games, and the motivations of the participants are understandable only if they are recognized as desire to engage in sport or play Business, in particular, Knight regarded as an outstanding game, possibly the greatest ever invented The ethical issues (not to gainsay the economic ones) that are involved here are as important as they are puzzling If games were merely a matter of recreation, one could simply classify this activity as one of the ways in which *Homo economicus* employs his time and the other resources of the economy, but the play element infuses much that is normally regarded as "work," and this raises a host of difficult questions concerning the ethics of competition, the nature of the nonzero-sum character of games, and the interpenetration of means and ends It even suggests that a competitive economy ought not to be viewed, as economists do, in instrumental terms, but as a manifestation or exemplification of human characteristics which are "fundamental" in a cultural sense In the competitive economy, man displays himself not only as *Homo economicus* and as *Homo faber*, but also as *Homo romanticus* and as *Homo ludens*

Politics too is a game, inseparable from the existence of social life Like other games it is subject to rules, but politics has a special significance in that it makes the rules by which all games (including itself) are played Knight repeatedly emphasized that the contraint of individual activity by rules does not in any way alter the essential fact that all social activity is the interaction of men, and all constraints upon individual action are decided by, and operated by, men, not by the abstraction of "law" It is only "romantic nonsense" he once remarked, to believe that one can have a "government of laws" that differs in any fundamental way from a "government of men " What matters in a polity is not the form of government, but how public opinion is generated To the extent that he would accept "democracy" as a label for a good political system, Knight adopted Lord Bryce's definition of it as "government by discussion " His frequent castigation of politicians was not intended to be a condemnation of politics, nor did it reflect any naive reverence for the automaticity of the market form of economic organization, the perfection of which he questioned more profoundly than any other economist, not exluding socialist ones His vexation with politicians was due to their tendency to turn "discussion" into "persuasion," which Knight regarded as coercion of a particularly insidious and dangerous type A good polity, a "democracy," is a society in which there is free exchange of ideas over the widest possible scope, unconstrained not only by censorship, but also by propaganda and guile In such a society, social policy will reflect a freely formed and eternally fluid public opinion which results from continuous discussion, most importantly of values rather than methods (which at bottom are really technical matters and can be settled by "science")

The operation of good political processes is very difficult, not only because of the tendency for "discussion" to degenerate into "persuasion," but because it is inherently difficult for any group of free men to arrive at collective decisions Arrow's "general possibility theorem" is not a serious difficulty for a democratic polity, unless one mistakenly identifies the process of voting, rather than discussing, as the essential feature of democracy Where discussion is free and extensive, individual preferences are modified and collapsed into a broad consensus, at least on fundamental values, which permits the machinery of representative or even plebiscitary processes to function effectively on specific issues Nevertheless, such processes of collective agreement are very difficult, and Knight perceived that it would grossly overload the machinery of democratic decision making to charge it with the task of deciding all matters, and all details, of collective cooperation A market economy does not require the performance of this practically impossible task Even a large society can be one of cooperation on the basis of exchange, which may take place on any terms which two parties agree is to their mutual advantage By means of the market mechanism a vast array of the practical problems of social existence are removed from the orbit in which coercion operates Any government which undertook to manage a modern economy in detail would have to be a comprehensive dictatorship, repressing not only voluntary exchange, but, in order to sustain itself, the free expression of opinion The view that a comprehensive government could be, or long remain, a protector of individual freedom is an error that rests upon denial of the elementary sociological fact that such a monopolistic center of power would inevitably attract to itself persons with a taste for coercing their fellow men The socialist labors under the delusion that while politicians and bureaucrats in a capitalistic society are venal, corrupt, and tyrannous, they will be paragons of honesty, justice, and self-constraint in a socialist one

The freedom that Knight regarded as the essence of liberal philosophy was not viewed by him, in purely negative terms, as the "absence of constraint " He differed from other liberals, like Hayek, in interpreting freedom as necessarily involving power, the ability to do what one wills This creates great difficulties in a world where what one may wish to do conflicts with the wants of others As a general philosophical principle, Knight accepted John Stuart Mill's proposition that one's legitimate freedom of action is constrained to the sphere within which there is no interference with that of others, but he recognized this to be a principle which leaves the largest problems of social life unsettled The economy, in particular, is a system in which conflicts of power are of the essence A perfectly competitive economy, operating upon the principle of voluntary exchange, assures that commutative justice will rule, no one is coerced to exchange against his interest But each person's power (and therefore his freedom) to engage in economic processes is a matter of the aggregate means at his disposal, and there is no reason to believe that the distribution of such means will be just in any commonsense or sophisticated sense of the term

This was, in Knight's view, the most conspicuous and most dangerous defect of the market economy He did not accept equality of income as a distributional ideal, but he viewed existing inequalities as gross and moreover, he felt that there were powerful forces at work in a private-property economy tending to increase this inequality it is easier for the rich to save than for the poor, the inheritance of property extends in time, and potentially without limit, the advantages gained by the successful accumulation of wealth at any point; the owners of property, unlike those whose persons are their only productive means, can operate in the market for assets as well as in the market for the services of assets, and so are more able, in a dynamic economy, to take advantage of opportunities for gain and to protect against losses Knight felt that the power of democratic government can be, and ought to be, used energentically against concentration of wealth (and power), but he was doubtful that it could do more than delay briefly the basic tendency He felt that the inevitable growth of inequality will erode the competitive market economy in itself, and it will also destroy its ethical justification

Knight was no romantic reactionary, looking back to some time in the past as the golden age of civilization, but neither was he a romantic progressive, viewing the evolution of society as tending inevitably toward the better and better He saw the history of the West as one of great human progress since the Renaissance The defeat of intellectual formalism, the rise of empirical science, the destruction of religious authoritarianism, and the effect of the market economy in (initially) dispersing power and weakening coercion were, in his view, the highlights of man's struggle to create a higher civilization But one cannot read Knight's essays without apprehending the deep pessimism of their author In part this was due to Knight's view that the competitive economy tends to economic inequality and therefore to self-destruction, in part it was due to his despondency over the qualities of man his perversity, prejudice, love of violence, and taste for evil, but more than these, he was dejected by the realization that the central problems of ethics were insoluble and that, even if that were not so, any society that devoted itself wholeheartedly to an ideal, even a good and noble one, would be a mean and servile place, stunting rather than enlarging man's capacity for civilization

Frank Knight disliked labels, his compound eye saw the weaknesses of any political or philosophical "position" at least as clearly as its strengths To classify him simply as a "liberal" would be misleading, and not only because that term has altered its meaning so greatly over the past century of its philosophical ascendancy in the West Even its older form was gravely flawed, in Knight's view, by its excessive individualism and failure to recognize adequately the social and the evolutionary character of man But in one great sense, which transcends these issues, he was indeed a liberal, in the tradition of Milton and Mill, believing that a good society is one that is uncertain of its truths and dedicates itself, not to an ideal, but to the eternal search

> The core of liberalism—what most distinguishes it from other views of life—is a manifold revolution in the conception of truth To say that belief is free is to say that truth is inherently "dynamic," subject to change and actually growing

> and changing The liberal interest in truth is one of curiosity and quest, not of mystical contemplation or adoration Any truth that is really "established" is no longer interesting, but a commonplace, even a bore Truth is the supreme example that liberal idealism looks at the values of life in terms of pursuit as well as possession Truth is an end when it is unknown or uncertain, and especially if controversial, hence the truth interest is finally a romantic one [*Freedom and Reform*, p 394]

Frank Knight was a liberal, but he was also a conservative, an idealist, and also an empiricist, a romantic, and also a utilitarian, and he was a great modern critic of all these His mental constitution was incapable of that monocular fixation which is the indispensable prerequisite of widespread fame or the founding of a "school " He leaves behind at his death many who have been deeply influenced by his thought, but no disciples—which is how he would have wanted it

Frank H. Knight as Economic Theorist

Arthur H. Leigh

Reed College

The greatness of Frank Knight as economic theorist rests as much on his contributions to the progress of economics on a broad front as to his more conspicuous theoretical breakthroughs, of which the theory of profit and the theory of capital are the most significant He defended the fundamentals of neoclassical economics by refining and rehabilitating them, not merely by reasserting them With explicit recognition of the depth and complexity of human motives, he abstracted from them the behavior patterns which are pertinent to economic analysis, without implying that actual behavior could be accurately predicted with those patterns alone [1] He demonstrated the effectiveness of a perfectly competitive market structure to achieve optimal allocative solutions but pointed out the difficulty of realizing or even defining with logical consistency the conditions of perfect competition [2]

Knight defined economics broadly as the study of social organization whose function is the allocation of given scarce resources to maximize the achievement of given ends He emphasized the universality of the choice problem imposed by scarcity and the generality of its solution in terms of marginal equalities As a writer and thinker of great breadth, he transcended disciplinary boundaries, including those of his own definition of economics, with profound philosophical discussions of value, freedom, justice, and their close links to economic theory

The Theory of Profit

In *Risk, Uncertainty, and Profit* (1921) Knight defined profit as a pure residual,[3] the amount of enterprise revenue remaining after the deduction of all enterprise costs Costs consist of payments at market prices for all productive services to which a contribution to the product can be imputed, including those provided by the owner-entrepreneur and his property They include "rents" for all exclusive advantages and privileges possessed by the enterprise which give rise to monopoly power, and payments for all measurable, and therefore insurable, risk The existence of a profit residual is therefore due to "uncertainty," which, in contrast to insurable risk, results from changes in market and technological conditions so unpredictable that their probability cannot be measured for the enterprise in actuarial terms[4] Generally anticipated change does not give rise to profit, for in this case competing entrepreneurs will bid up the prices of inputs and bid down the prices of product even before the change actually occurs, maintaining equality between revenue and cost The receipt of profit is temporary, lasting only until the innovation or adaptation which gave rise to it has been generally imitated[5] The realization of positive profit rewards the entrepreneur who anticipates change sooner and more accurately than do the majority of competing entrepreneurs The average magnitude of profit in the aggregate tends to be low if entrepreneurs are generally optimistic about their prospects and bid up the prices of productive services Knight suggests reasons for believing that aggregate profit may be negative with average net losses cutting into other (imputed) income People are commonly willing to gamble when the odds are known to be stacked against them Businessmen tend to be overly optimistic about their own capacities and chances for success People respond to the dynamic stimulus to outbid their rivals and play the game, and the prestige and independence of being one's own boss has wide appeal

---

[1] Cf Buchanan 1968, p 426, Bose 1970, pp 7–10, Breit and Ransom 1967, pp 198–200, Rima 1972, pp 289–91

[2] A noteworthy example is his famous controversy with Pigou on the effects of increasing returns on the optimality of the competitive solution and on the relationship between social and private costs (Knight 1936*a*, chap 8, pp 217–36, cf Rima 1972, pp 288–89, 294–95, and Breit and Ransom 1967, pp 191–200)

[3] Knight acknowledges prior suggestion of his theory of profit by F B Hawley (Knight 1921, pp 31, 41–46) and the nineteenth-century German economists von Thunen and von Mangholt (Knight 1921, p 26).

The entrepreneurial function is that of making decisions for the successful adaptation of the enterprise to change, and in a progressive society, acting as the economic pioneer initiating useful innovations The entrepreneur is a specialist in bearing uncertainty, including the uncertain consequences of his own decisions He guarantees fixed contractual payments to the contributors of productive services, thus relieving them of uncertainty It is the function of profit (or the anticipation of profit) to motivate and reward this socially desirable behavior [6]

---

[4] For critical commentaries on Knight's distinction between risk and uncertainty and its relation to probability theory and profit, see Hicks (1931), pp 170–71, 175, Arrow (1951), pp 408, 415–17, 426, and Fellner (1965), pp 29–31 Hicks credits Knight with placing the theory of profit on a firm foundation—the dependence of profit on uncertainty but argues that a satisfactory theory of profit can be developed without Knight's concept of true "uncertainty" or nonmeasurable risk Arrow and Fellner agree that Knight failed to develop a clear theory of uncertainty as distinct from risk, but Arrow characterizes as an 'analytic point of considerable interest" Knight's assertion that if all risks were measurable, then risk aversion would not give rise to any profit (p 427)

[5] Schumpeter, like Knight, held a dynamic residual theory of profit, but he emphasized innovation and development rather than uncertainty and unpredictable change as its determinants (cf Schumpeter 1934, pp 128–36, 172, 1939, pp 102–9)

Before 1921, the profit concept was, with few exceptions, a variety of confusing mixtures composed of return on capital, wages of management, premia for risk, monopoly rent, entrepreneurial income, and exploitation of workers Knight isolated profit as a distinct distributive share and identified its essential social function together with that of the entrepreneur. In the process of doing so, he reexamined and improved the theories of competition, product imputation, knowledge and uncertainty, and enterprise organization He helped to discredit the common belief that "profit" can be indiscriminately taxed or bargained away with impunity The modern technical and textbook literature reflects his theory of profit significantly, but not completely Nearly all distinguish profit from interest, risk premia, and managerial wages Fewer distinguish clearly between profit and monopoly rents "Profit" in business, tax, accounting, and popular usage is still the old confusing mixture

The Theory of Capital and Interest

Knight developed his theory of capital and interest in a series of controversial journal articles and essays, the first of which appeared in 1916 and the remainder between 1931 and 1944 [7] It is best understood as the core of his theory of production and growth and as a reaction against important elements of Austrian,[8] classical, and neoclassical economics He rejected the then widely accepted "Austrian" capital theory, according to which capital is intermediate goods first produced by "original" productive agents such as labor and natural resources and then used up in the production of final goods over a definable period of time In Austrian theory, the quantity of capital used varies directly with, and is measured by, the average length of the time period between the investment of original factor services and the consumption of the final product The productivity of a process increases with greater "roundaboutness" (capital intensity), but the lengthening of the period of production is subject to diminishing returns A general preference for present over future goods of like kind and quantity, expressed as a rate of discount, means that savers must be compensated for "waiting" and limits the extension of the period of production The equilibrium rate of interest for the economy is established by the relationship between an upward-sloping supply curve for saving based upon time preference and a downward-sloping demand curve for capital based upon the marginal productivity of increased roundaboutness

---

[6] The social function of profit as Knight saw it was most clearly spelled out in an article which appeared much later than *Risk, Uncertainty, and Profit* (Knight 1942)

[7] Knight *1916, 1931, 1933a, 1933b, 1934b, 1935a, 1935b, 1935c, 1936b, 1937, 1938*, 1941, and 1944

[8] Writers with whom Knight took issue were Bohm-Bawerk, Menger, Wieser, and Jevons of the earlier generation, and included Mises, Kaldor, and Hayek among his contemporaries In many important respects, Knight's concepts of capital and the productive process follow those of J B Clark (cf Clark 1914, chaps 9, 10, 18, and 20)

In contrast to these views, Knight defined production as the rendering of want-satisfying services by a large variety of coordinate productive agents and the creation of net additions to the stock of productive agents With production in the first sense, consumption (the enjoyment of services) is simultaneous With production in the second sense, no consumption corresponds unless or until society as a whole uses up its stock of productive capacity, a "decadent" condition not characteristic of modern societies [9] There is therefore no identifiable, much less measurable, period of production short of the whole interval between the historic origins of the economic system and its ultimate termination

Useful classification of "productive agents" requires distinctive categories whose component units are sufficiently interchangeable to permit the correct application of the marginal productivity principle The traditional land, laborers, and capital goods trilogy is therefore more misleading than useful Knight's "capital" is not a category of productive agents classified by origin or physical characteristics, but the present value of all nonhuman productive agents which yield a flow of valuable services through future time [10] For purposes of theory and accounting, this flow is reduced to a perpetuity by providing for the maintenance or replacement of concrete productive agents out of their product flow This replacement process makes up much of what is commonly called "production " The stock of capital is thus the value embodied in a vast variety of objects of every degree of durability At the indistinct borderlines of the concept are intangibles (such as knowledge) which may or may not be appropriable, and free human beings whose productive capacity is partly the result of economically motivated investment but for legal reasons is not subject to capitalization as property

To measure a quantity of existing capital, the present value of a unit of productive capacity must be found by capitalizing the yield imputable to it at some given rate of interest The appropriate rate is the ratio between an increment of perpetual annual service yield and the minimum cost, in terms of foregone consumption, of constructing an increment of productive capacity capable of yielding it, and is determined at the economy's margin of investment.[11] Free competitive markets establish determinate values for both the yield and the cost of an investment and tend, in spite of uncertainty, to bring rates of yield on all lines of investment into mutual equality. This is the marginal rate of return on investment for the economy which is used to capitalize the yield of all existing nonhuman productive agents and thus to measure the capital embodied in them.

---

[9] This does not mean, of course, that no additional consumption corresponds with (results from) the additional net output produced by the net additions to the stock of productive agents, once they are in operation But that additional consumption cannot at the same time be attributed to production of services by the new agents and the production of those new agents by previously existing agents

[10] Cf Schumpeter 1954, pp 900, 906, and 1018

Knight argued that this rate, and the loan rate which tends to conform to it, are determined independently of any supply function for savings or loanable funds. The short-run demand function for liquid capital is infinitely elastic because the marginal productivity of capital is a function of the total stock of capital, and investment over a period as short as 1 year adds too little to that stock to affect its marginal productivity. The choice of savers between more current consumption and larger streams of future income therefore affects not the rate of interest but the quantity of saving and investment. Knight considered the Austrian attempt to explain this choice in terms of time preference to be misspecified. Properly stated, the question is whether, *ceteris paribus*, people would generally prefer to receive a stream of income of given duration and total magnitude in increasing, equal, or decreasing installments, apart from the opportunity to invest and earn interest on the earlier installments. Common observation suggests that the first option would be at least as popular as the third, a result which would run contrary to the time-preference theory.

Knight assumed that the economy's marginal productivity of investment is a decreasing function of the stock of capital and speculated that the elasticity of the function approximates unity. The rate of return could never reach zero unless all services which capital goods could render became free goods. But it was not until late in the development of his theory of capital (1944) that he explained diminishing returns to investment.[12] The comprehensiveness of his capital concept posed a problem. A growth model in which all agents are freely augmentable by investment appears to afford no basis for diminishing returns to investment in general, for there are no agents fixed in quantity to which successive increments of variable agents can be applied with consequent successively diminishing increments of product. As solution, he postulated the existence of some productive agents which could be augmented only at sharply increasing cost in terms of other agents, with absolute fixity of supply as the limiting case There are echoes here of the traditional land versus capital distinction, but Knight does not identify his fixed agents with natural resources He saw no clear historical trend in the rate of interest and denied the existence of a long-run equilibrium rate Discovery and invention, themselves a form of investment, together with population growth, have tended to offset the downward influence of capital accumulation

---

[11] Knight expressed his definition of the net rate of return on investment in algebraic form as follows: $SA^L(A^c - 1) = R(A^L - 1)$, where S is the number of dollars invested per year over a construction period of c years, R is the imputed annual yield of the asset in dollars, L *is the life of the asset in years,* A *equals* $(1 + i)$, i *is the rate of return* (the unknown). Note that i is a net (perpetual) rate of return even though the investment on which the equation is based is assumed to have limited life, and that the rate enters into the cost side (as well as yield side) of the equation whenever the construction of the asset requires an interval of time (see Knight 1956, pp. 49–50, 1934*b*, pp. 265–67, 1936*b*, pp. 445–46).

[12] Knight (1944). In an earlier article, Knight (1938) offered an explanation of diminishing returns to capital in terms of diminishing marginal utility of consumption services, but later rejected it in favor of the explanation which appeared in the 1944 article.

Knight succeeded in freeing capital theory from preoccupation with the characteristics of particular productive agents and an array of distracting distinctions and classifications His theory of capital and interest is a long-term, aggregative model of production and growth emphasizing the interdependence and continuity of the productive system, the measurement and productivity of capital, and the cost and consequences of accumulation States of transition and disequilibrium are treated only incidentally, though it is clear that Knight considered the durability and transferability of concrete productive agents to be important as influencing the adaptability of the economy to changing conditions His subordination of disequilibrium analysis may have been partly responsible for his overemphasis upon demand at the expense of supply in interest-rate determination An infinitely elastic short-run demand function for investment funds implies that within an interval as short as 1 year equally profitable investment opportunities are unlimited, that the relative prices of investment goods do not rise in response to high investment activity, and that even in a continuously changing economy, prospective rates of return on all particular lines of investment remain equal in the short run [13] This is not a realistic set of assumptions [14]

The Concept of "Cost"

In the realm of conceptual refinements, Knight's view of "cost" occupies a prominent position The cost of anything is the value of the best alternative choice foregone to obtain it, not the "pain" or disutility of working or waiting endured in production or investment The contribution was not the opportunity-cost idea in itself, which had been emphasized by the Austrian economists in the 1870s, but in clearing away the pain-cost notion which the Austrians had failed to do This simplification of the cost concept facilitated another of Knight's contributions, a clarification of the relationships among cost, utility, value, and price within a general interdependent system It directs attention to the choice problem central to efficient resource allocation and away from moralistic justification of the prices paid for productive services in terms of subjective sacrifice In a market economy, income distribution is a matter of the pricing of productive services and the ownership of productive capacity Distributive justice, with which Knight was deeply concerned, was to him a matter of free market competition and (more important as problems unsolved) inequality in inherited capacities and wealth, and in the opportunity to acquire, develop, and market valuable productive services By denying the usefulness of a "pain" theory of cost, Knight did not mean that all choices are among pleasurable alternatives The foregone alternative can as well be the avoidance of fatigue, boredom, or waiting as the enjoyment of leisure or immediate consumption satisfaction Rational choice does not depend upon the exact location of the zero point between utility and disutility

---

[13] Knight's own qualification of his demand theory of interest rates was that since investment planning requires time, sudden changes in saving could temporarily glut or starve the market for loanable funds and affect the interest rate (Knight 1934*b*, pp 285-86)

[14] For a highly critical commentary on Knight's theory of capital, see Seligman (1962), pp 646-65

Conclusion

I conclude with emphasis upon the breadth of Knight's contributions to the improvement of economic theory As one of the last great general economic theorists to rely entirely upon nonmathematical exposition,[15] he was keenly aware of the weaknesses of the conceptual categories which he inherited from his predecessors Emphasizing the fact that all scientific analysis is based upon and limited by the use of distinct classes of relevantly homogeneous phenomena, he rejected or subjected to critical scrutiny such traditionally accepted categories as land, labor, capital, profit, interest, production, consumption, cost, value, competition, risk, the entrepreneur, and economics His lifelong effort to increase the usefulness and the logical clarity of the basic concepts of the field led him frequently to qualify his own concepts, elaborate upon their meaning and limitations, and digress from the central line of argument This tendency made many of his writings discouragingly complex to some readers, but it contributed substantially to the intensely stimulating quality of his work It was part of his effort to get to the fundamentals of the discipline, to construct firm foundations, usefully related to the real world as we perceive it, on which to construct our increasingly elaborate theoretical superstructures, and to set economic theory in the direction of discovering and elucidating rather than obscuring the "truth"

---

[15] Having begun his intellectual life as a physicist, Knight's choice of verbal exposition was a matter of preference, not necessity

References

Arrow, Kenneth J "Alternative Approaches to the Theory of Choice in Risk-taking Situations" *Econometrica* 19 (October 1951) 404–37

Bose, Ashalata *The Methodology and Scope of Economics* Calcutta Bookland Private, 1970

Breit, William, and Ransom, Roger L *The Academic Scribblers* New York Holt, Rinehart & Winston, 1967

Buchanan, James M "Knight, Frank H" In *International Encyclopedia of Social Sciences*, edited by David Sills Vol 8 New York Macmillan, 1968

Clark, John Bates *The Distribution of Wealth* New York Macmillan, 1914

Fellner, William *Probability and Profit*, Homewood, Ill Irwin, 1965

Hicks, John R "The Theory of Uncertainty and Profit" *Economica* 11 (May 1931) 170 89

Knight, Frank H "Neglected Factors in the Problem of Normal Interest" *Q J E* 30, no 2 (February 1916) 279 310

——— *Risk, Uncertainty and Profit* Boston Houghton Mifflin, 1921

——— "Professor Fisher's Interest Theory A Case in Point" *J P E* 39 (April 1931) 176–212

——— "Interest" *Encyclopedia of the Social Sciences* Vol 8 New York Macmillan, 1933 (*a*)

——— 'Capitalistic Production, Time, and the Rate of Return" In *Economic Essays in Honour of Gustav Cassel* London Allen & Unwin 1933 (*b*)

——— "Profit" *Encyclopedia of the Social Sciences* Vol 12 New York Macmillan, 1934 (*a*)

——— "Capital, Time, and the Interest Rate" *Economica* n s 1 (August 1934) 257–86 (*b*)

——— "Comment" *J P E* 43 (October/December 1935) 625–27 (*a*)

——— "Professor Hayek and the Theory of Investment" *Econ J* 45 (March 1935) 77–94 (*b*)

——— 'The Theory of Investment Once More Mr Boulding and the Austrians" *Q J E* 50 (1935) 36–67 (*c*)

——— *Ethics of Competition* 2d ed New York Harper, 1936 (*a*)

——— "The Quantity of Capital and the Rate of Interest" *J P E* 44 (August and October 1936) 433–63, 612–42 (*b*)

——— "A Note on Dr Lange's Interest Theory" *Rev Econ Studies* 4 (June 1937) 223–30

——— "On the Theory of Capital In Reply to Mr Kaldor" *Econometrica* 6 (January 1938) 63–82

——— Professor Mises and the Theory of Capital" *Economica* n s 8 (November 1941) 409–27

——— 'Profit and the Entrepreneurial Function" *J Econ Hist* 2 (December 1942) 126–32

——— "Diminishing Returns from Investment" *J P E* 52 (March 1944) 26–47

——— *On the History and Method of Economics* Berkeley Univ California Press, 1956

Rima, I H *Development of Economic Analysis* Rev ed Homewood, Ill · Irwin, 1972
Schumpeter, Joseph A *The Theory of Economic Development.* Cambridge, Mass Harvard Univ Press, 1934
———. *Business Cycles* Vol 1 New York McGraw-Hill, 1939
———. *History of Economic Analysis* New York Oxford Univ Press, 1954.
Seligman, Ben "F Knight and Abstractionism " In *Main Currents in Modern Economics* Glencoe, Ill Free Press, 1962

Communications

Rational Response to the Money Supply

Richard Roll

European Institute for Advanced Study in Management

I. Introduction

Knowledgeable consumers should be attentive to published measures of the money supply. They have an incentive to forecast money's future quantity as accurately as possible, for they must decide which commodities, assets, and labor income should be contracted now, in nominal terms, and which should be left for purchase or sale at future (unknown) prices.

Information about the money supply has a central role in consumption forecasting and choosing because it is commonly and correctly believed to be the most important influence on commodity prices. By forecasting the money supply into the future and estimating its effect on the sequence of commodity prices, the consumer is performing a rational decision calculation, identical in substance with that performed by an investor estimating the effect of predicted future earnings on the sequence of a firm's common stock price.

By the very act of making decisions on the basis of anticipated money stocks, however, consumers decrease the empirical connection between future changes in money and the current price of goods. The efficient commodity market would display a minimal connection between the two because commodity prices would be bid quickly to a discounted level related to the predictable portion of the monetary sequence. Again

This research was supported by a Ford Foundation Fellowship. However, the conclusions, opinions, and other statements are mine alone and are not necessarily those of the foundation. The participants of the NBER conference on Rational Expectations and Time Series Analysis at the University of Minnesota in March 1973 provided many useful comments. In particular, I wish to thank William Brainard, William Brock, Edmund Phelps, and Karl Shell (without suggesting that they are, in any way, responsible for the results).

the analogy to a perfect stock market can be made a firm whose earnings have grown period after period will not find its common stock changing in price after the next earnings increase unless those earnings deviate from the anticipated

Unlike corporate earnings, the supply of fiat money represents virtually no real resource It can be manipulated willfully and costlessly, and rules for the distribution of new nominal money balances can be chosen arbitrarily by the fiat issuer The current distribution rule in the United States involves direct exchange of new money for real goods and services and financial assets, but this is certainly not the only possible rule or even the rule which is most frequently discussed by monetary theorists They generally suppose that the distribution rule is unimportant and indeed has no bearing whatsoever on the pattern of commodity prices observed in the economy [1]

In this paper, I hope to argue well for a contrary hypothesis The nominal money distribution rule is of crucial significance provided that consumers behave by adjusting their current portfolios of money, assets, and consumption on the basis of the anticipated future monetary sequence Although the rules to be discussed here are primitive, they do illustrate the consequences, for monetary policy, of rational consumer behavior Hence, they may claim to have a prominence beyond their reality

II. Asset, Money, and Commodity Demand

On the most disaggregated level, the individual consumer can be envisioned as selecting a feasible consumption program over his lifetime In the initial period, his choice set contains probability distributions only and does not include a guaranteed future level of consumption The time-honored method of converting this choice set into a set containing recognizable economic quantities is to suppose he maximizes the expected value of a utility function defined over current and future consumption levels and then to relate such levels to an initial wealth endowment and to quantities of assets selected currently with the purpose of storing wealth for the future

Throughout the paper, all costs attendant to buying and selling assets and differential tax effects will be neglected Within these restrictions, the consumer's hope is to maximize $E[U(C_1, \tilde{C}_2, \quad \tilde{C}_n)]$ while remaining within the constraints imposed by

$$W - M_1 - p_1 A_1 - P_1 C_1 = 0 \qquad (1)$$

$$\tilde{P}_j \tilde{C}_j - \tilde{y}_j - (\tilde{M}_{j-1} - \tilde{M}_j) - \tilde{p}_j(\tilde{A}_{j-1} - \tilde{A}_j) = 0, \; j = 2, \quad n - 1 \qquad (2)$$

$$\tilde{P}_n \tilde{C}_n - \tilde{y}_n - \tilde{M}_{n-1} - \tilde{p}_n \tilde{A}_{n-1} = 0 \qquad (3)$$

---

[1] The quantity of nominal money balances is emphasized, of course It is only the distribution rule which is considered unimportant

where U = utility, E = mathematical expectation, C_j = goods consumed in period j, P_j = consumption goods price in period j, A_j = quantity of bonds (variable price asset) held from j to $j + 1$, p_j = bond price in period j, M_j = quantity of money[2] held from period j to $j + 1$ (nominal balances), W = initial wealth (human and nonhuman capital),[3] y_j = labor income earned in period j (this is assumed to be unrelated functionally to asset and money holdings), n = maximum possible life span, $\sim$ denotes random variables as of the first (current) period[4] A solution to this problem requires the "stochastic calculus" of Kushner (1965) because the constraints (2)–(3) contain stochastic elements Kushner's technique has been discussed in detail in economic applications by Long (1971), and the reader should refer to that paper or to Kushner's own for more information Here, only the results (first-order extremum conditions) for the solution to the above problem are given For the asset money, they are[5]

$$E(\tilde{U}_1) = E[\tilde{U}_2/(1 + \tilde{I}_2)]$$

$$\tilde{U}_j = \tilde{U}_{j+1}/(1 + \tilde{I}_{j+1}), j = 2, \quad , n - 1 \tag{4}$$

where subscripted U's indicate partial derivatives of the utility function and I_j is the rate of goods price inflation between periods $j - 1$ and j, $I_j \equiv (P_j - P_{j-1})/P_{j-1}$

The optimum conditions (4) imply rather simple marginal trade-offs between time preferences and commodity price changes When one realizes that U_j/U_{j+1} has the units of a real interest rate (i e , a commodity interest rate), plus unity, the conditions are seen to be the monetary asset's equivalent to the Irving Fisher (1930) relation between real interest rates, nominal interest rates, and the rate of inflation (For money, the nominal rate of interest is exactly zero)

There should be as many different U_j's and I_j's in each period, as there are different consumption goods By not denoting the quantities in (4) like "$U_{j,i}$" and "$I_{j,i}$" for the marginal utility and inflation rate corresponding to good i, I have assumed implicitly a single composite consumption good This has been done solely for notational ease Explicitly accounting for k different available consumption goods in each period would place k equations in (4) for every one there now. This would cause no special difficulty aside from the inevitable problem of selecting the "correct" commodity price index for empirical work

---

[2] Money is defined as assets whose nominal values are always fixed, e g , currency, coin, non-interest-paying demand deposits Money is assumed to bring no nonpecuniary utility

[3] For the purpose of this illustration, constraints on the nonmarketability of (human) capital are ignored

[4] Stochastic prices and labor incomes are realized at the beginning of the period denoted by their subscript For example, p_3 is the actual bond price exactly at the start of period 3, two full periods after the beginning of the process Goods purchases and consumption are assumed to occur during each period after prices and income are observed Repurchases of assets and money also are assumed to occur during each period at the prices observed at its beginning

[5] The equations in (4) are obtained by combining several first-order conditions to eliminate Lagrange multipliers There are also conditions on asset holding which derive from the maximization problem but they will not be discussed in this paper

For $j \geqq 2$, note the relation in (4) of several stochastic quantities whose equilibration is not under the investor's current control The sense of these equations relates to consumption decisions upcoming in future periods As they must depend on previous outcomes, the elements of these expressions are obviously stochastic now, but as the investor will choose optimally for any outcomes whatever, the equalities themselves are certain They are, in fact, the "stochastic calculus" analog to the dynamic programming assumption that all future decisions will be made optimally conditional on the states that have occurred up until the decision period

The equations in (4) may seem to imply that uncertainty exists only in period 1 This is not the case At the end of period 1, asset and consumption prices p_2 and P_2 are known and the consumer makes a second round of choices which brings him to the following positions

$$E(\tilde{U}_1) = E(\tilde{U}_2)/(1 + I_2),$$
$$E(\tilde{U}_2) = E[\tilde{U}_3/(1 + \tilde{I}_2)],$$
$$\tilde{U}_j = \tilde{U}_{j+1}/(1 + \tilde{I}_{j+1}),\ j = 3, \quad , n - 1$$

These differ from the initial first-order conditions because I_2 now is nonstochastic, marginal utilities $\tilde{U}_1$ and $\tilde{U}_2$ may still be stochastic even though consumption has occurred because the marginal utility of consumption in an early period may be a function of (stochastic) consumption in later periods For additive utility functions, however, the first equation above would vanish because U_1 would not depend upon later choices

III. Simple Models of Nominal Money Supply

Helicopters and Other Mints

The lucidness of Friedman's (1969) hypothetical economy recommends its adoption in the following analysis Friedman supposes a helicopter bearing fiat money to circle over the economy dispensing (or collecting) bills at given rates No resources are necessary for these flights nor for producing the added (or decremented) currency Friedman's purpose was to discover the socially optimum frequency of flights, a topic which depends a great deal on the value of money's specific nonpecuniary services [6] My purpose, however, is to examine the effect of such flights on current commodity prices (and thus on anticipated inflation rates), and for this the nonpecuniary qualities of money simply get in the way, adding extra terms to the derived equations but no extra substance to the economics Consequently, all nonpecuniary qualities of money except its perfect nominal safety will be ignored here Of course, Friedman's essay also discussed the process of inflation, but in its concentration on the major issue of optimum money supply, it neglected the rules used by the helicopter's bombardier to distribute money among individuals

---

[6] See Meltzer (1970) for a clear enunciation of this point

Consider the following three mutually exclusive simple money supply rules New money is distributed (or collected) (*a*) proportionately to money balances held in the first period;[7] (*b*) proportionately to money balances held *at the beginning of each period*, (*c*) randomly (and not proportionately to money balances) These rules are so simple that we can just append them to constraints in each individual's demand problem, to equations (1), (2), or (3) of the last section With rule *a*, money is distributed during all future periods in amounts proportional to M_1, the quantity of money balances held at the beginning of period 1 The individual still intends to maximize expected utility of consumption, $E[U(C_1, \tilde{C}_2, \quad , \tilde{C}_n)]$, and his first-period wealth constraint *is unchanged* But under the assumption of a growth in money at the rate α, distributed in each future period proportionately to first-period holdings, his second- and later-period constraints become[8]

$$\tilde{P}_j\tilde{C}_j - \tilde{y}_j - (\tilde{M}_{j-1} - \tilde{M}_j) - \tilde{p}_j(\tilde{A}_{j-1} - \tilde{A}_j)$$
$$- M_1[e^{(j-1)\alpha} - e^{(j-2)\alpha}] = 0, j = 2, \quad , n - 1 \qquad (2a)$$

$$\tilde{P}_n\tilde{C}_n - \tilde{y}_n - \tilde{M}_{n-1} - \tilde{p}_n\tilde{A}_{n-1} - M_1[e^{(n-1)\alpha} - e^{(n-2)\alpha}] = 0 \qquad (3a)$$

A straightforward application of the stochastic calculus to this problem will provide a different set of first-order extremum conditions than that

---

[7] This is the rule used in Friedman's simplest model (1969, p 9)

[8] Since, the growth rate of money is α, we have

Beginning of Period	Total Money Supply, M^s	Increment in Supply from Last Period
1	$M_1^s = M_1^s$	
2	$M_2^s = M_1^s e^\alpha$	$M_1^s(e^\alpha - 1)$
3	$M_3^s = M_1^s e^{2\alpha}$	$M_1^s(e^{2\alpha} - e^\alpha)$
j	$M_j^s = M_1^s e^{j\alpha}$	$M_1^s[e^{(j-1)\alpha} - e^{(j-2)\alpha}]$

In period 2, the investor receives $M_1(e^\alpha - 1)$ from the helicopter and he has held M_1 from period 1 Thus, his total money balance is $M_1 + M_1(e^\alpha - 1) = M_1e^\alpha$ at the beginning of period 2 He chooses to carry the amount M_2 over from period 2 to 3 Thus his money balance in period 3 is M_2 plus the increment he receives from the helicopter, $M_1(e^{2\alpha} - e^\alpha)$, which is, by the assumed rule, based on first-period balances, etc , for later periods formerly obtained with a fixed money supply. Corresponding to equation (4), which depicted the (fixed-supply) marginal trade-off between commodity price inflation and time preference, one will now find

$$E(\bar{U}_1) = E[\bar{U}_2/(1 + \bar{I}_2)]e^{(n-1)\alpha},$$

$$\bar{U}_j = \bar{U}_{j+1}/(1 + \bar{I}_{j+1}),\ j > 1 \tag{4a}$$

For the first-period equation, (4a) indicates a rate of commodity price inflation $n - 1$ times the rate of monetary expansion.[9] For later periods, however, the inflation rate is uninfluenced by money's growth even though the two events occur simultaneously. Given neutral time preference by all consumers, these conditions actually imply a zero rate of change in commodity prices between adjacent periods beyond the first two, which implies that money supply rule *a*, linking all future distributions to first-period balances, must have its entire effect during the first period. The helicopter's promised appearance sets off a clamor for period 1 money balances that only ceases when period 1 prices have been bid down (i.e., money's period 1 commodity price has been bid up), to the point that an inflation rate of $e^{(n-1)\alpha}$ is anticipated between periods 1 and 2. The inflation rate between the first two periods is $n - 1$ times the rate of money expansion then because new money to be dispensed later has been discounted in advance, thus raising even further the benefit of holding money in period 1. Under rule *a*, money balances in period 2 are not linked in any way to future distributions of new money so there is no incentive to increase the demand for them similar to the incentive existing for period 1 balances.

Since this conclusion is at variance with Friedman's,[10] it should be worth considering the assumptive differences which must have been responsible. I believe they can be reduced to a single item. Friedman's individuals are assumed to be either myopic or irrational. His "individual is not able to affect the amount of additional cash he receives by altering the amount of cash he holds" (in the first or in any other period) (1969, p. 9). With this presumption, there is no reason for an individual to forecast future helicopter flights or to perform any portfolio adjustments that would ordinarily bring gains in expected utility. If the individual were not so restricted, he would find it advantageous to increase money balances in period 1 as I have assumed he actually does. It must be admitted, however, that the hypothetical individual of this paper is on the opposite extreme of the myopia spectrum from Friedman's individual; here, he is presumed to realize the helicopter is coming and to understand its distribution rule perfectly.

---

[9] With this rule and with the others to follow, the individual first-order optimum relations sometimes are discussed as if they were market equilibrium conditions. This is strictly valid only if all investors have common subjective probability beliefs, i.e., homogeneous expectations. In that case, ratios of marginal utilities between adjacent periods, (U_j/U_{j+1}), must be the same for all investors in equilibrium because individual optimum *conditions such as* (4a) *bind the marginal utility ratios to (commonly held) assessments* of future inflation and future money growth. With nonhomogeneous expectations, expressions with the same form as (4a) would be in effect as market-clearing relations, but *the* $\bar{U}_j$'s *would* have *to be interpreted as complex weighted averages of individual* marginal utilities.

[10] He concludes that a continual monetary expansion or contraction at the rate e^{α} will bring, after initial short-term adjustments, a continual inflation or deflation at the same rate.

Given the monetarily prescient consumer, a most interesting feature of rule *b*, which distributes new money proportionately to balances held at the beginning of each period, is an equivalent market phenomenon to that Friedman has found using rule *a* and a myopic or irrational individual. To prove this, notice that constraints (2) and (3) of the basic maximization problem have different forms under rule *b*. Money is now distributed at a growth rate α to holder of balances at the beginning of each period so that notationally, the constraints for future periods are

$$\tilde{P}_j\tilde{C}_j - \tilde{y}_j - e^{\alpha}\tilde{M}_{j-1} + M_j - \tilde{p}_j(\tilde{A}_{j-1} - \tilde{A}_j) = 0, j = 2, \ldots, n-1, \tag{2b}$$

which is the same as under rule *a* only for period $j = 2$, and

$$\tilde{P}_n\tilde{C}_n - \tilde{y}_n - e^{\alpha}\tilde{M}_{n-1} - \tilde{p}_n\tilde{A}_{n-1} \tag{3b}$$

After performing a few stochastic calculus operations, one now will find the process of inflation described by

$$E(\tilde{U}_1) = E[\tilde{U}_2/(1 + \tilde{I}_2)]e^{\alpha}$$
$$\tilde{U}_j = [\tilde{U}_{j+1}/(1 + \tilde{I}_{j+1})]e^{\alpha}, j = 2, \ldots, n-1, \tag{4b}$$

which, under certainty and with neutral time preference, would imply an inflation rate equal to the money growth rate in every period.[11]

Now consider the third rule, *c*, whereby new fiat money is distributed without regard to initial money balances. Suppose $\tilde{M}_j^s$ is the (random) number of bills received by an individual during period j. If these are truly distributed at random without hope of being affected through resource expenditures in earlier periods, future constraints on the individual's maximization problem are altered to become

---

[11] Under rule *b* it is also easy to confirm the Fisherian relation between nominal interest rates, real (commodity) interest rates, and the anticipated rate of inflation. The cause of an inflation premium in the nominal interest rate is quite removed from Fisher's (1930) explanation, however. The helicopter distributes bills to money holders only and the inflation premium is a penalty demanded by bond-holders to compensate for their not benefiting by the helicopter's munificence. Since the holder of a quantity M_j of money at the beginning of period j receives $M_j(e^{\alpha} - 1)$ from the helicopter during j, he could only be induced to hold bonds if they were expected to appreciate in nominal value by at least the same relative amount.

$$\tilde{P}_j\tilde{C}_j - \tilde{y}_j - (\tilde{M}_{j-1} - \tilde{M}_j) + \tilde{M}_j^s - \tilde{p}_j(\tilde{A}_{j-1} - \tilde{A}_j) = 0,$$

$$j = 2, \quad . , n - 1 \tag{2c}$$

$$\tilde{P}_n\tilde{C}_n - \tilde{y}_n - \tilde{M}_{n-1} + \tilde{M}_n^s - \tilde{p}_n\tilde{A}_{n-1} \tag{3c}$$

The first-order extremum conditions are exactly the same as those of Section II where the money supply was fixed This is because $\tilde{M}_j^s$ is not a decision variable It is not under the individual's control and thus cannot be chosen to achieve an optimum asset portfolio or an optimum consumption sequence Of course, this does not imply a total lack of effect by M^s on sequences of asset and commodity prices The basic marginal condition relating to inflation rates will still be equations (4), but these relations arise by equating certain random Lagrange multipliers that will be altered by the additions of more money in the future [12] In fact, one may definitely infer a change in marginal time preference between the equilibrium conditions under absolutely fixed money supply and under rule *c* variation in supply If investors anticipate future random money increases, they will attempt to decrease current money balances by exchanging them for assets or commodities This will have at least two effects It will increase current commodity prices, P_1, and it will decrease the first marginal utility, U_1, to the extent that risk averters use some newly purchased commodities for period 1 consumption.[13] Starting from a position of neutral time preference, $(U_1 = U_j)$, which implies zero inflation,[14] equilibrium first-order conditions will move to a position

---

[12] Specifically, the first-order conditions responsible for equations (4), with, for example, $n = 2$, are

First-Order Condition	For Decision Variable
$E(\tilde{U}_1 - \tilde{J}_1P_1) = 0$	C_1
$E(\tilde{J}_1 + \tilde{J}_2) = 0$	M_1
$\tilde{U}_2 + \tilde{J}_2\tilde{P}_2 = 0$	$\tilde{C}_2$

where the J's are random Lagrange multipliers J_1 corresponds to the wealth constraint for period 1, $W = P_1C_1 + M_1 + p_1A_1$, and J_2 to the constraint for period 2, $\tilde{P}_2\tilde{C}_2 = M_1 + \tilde{P}_2(\tilde{A}_1) + \tilde{M}_2^s + \tilde{y}_2$ The constraints are, of course, first-order conditions with respect to $\tilde{J}_1$ and $\tilde{J}_2$ Since period 2's constraint is altered by the addition of $\tilde{M}_2^s$, its associated Lagrangian, J_2, will also be altered, and since $(\tilde{U}_2/\tilde{P}_2) = -J_2$, the first-order condition for $\tilde{C}_2$, and all other first-order conditions involving J_2, will also be changed

[13] This will be mitigated somewhat by consumers carrying over more commodity inventory to future periods

[14] It should be emphasized again that the implication of zero inflation (brought about by an assumption of neutral time preference) occurs only if money's nonpecuniary benefits are ignored

of positive inflation, $P_j > P_1$, and negative marginal time preference, $U_j > U_1$, such that equations (4) still hold Note again the effect of advanced discounting—future money issues change current prices, P_1, as well as future prices, P_j, even though, in this case, the extent of future price changes will be larger than the current price change

Under a random distribution rule, therefore, the inflation sequence could possibly match the sequence under rule *b* a continual price change at the rate of monetary expansion This is not necessary, however, because consumers have the freedom to purchase and hold assets and commodity inventories rather than money By purchasing inventories initially, for example, just when the helicopter flights and random mintings are first announced, consumers would cause a much larger price change initially than the monetary expansion in period 1 To offset this, later inflation would have to evolve at a rate lower than the rate of money growth If asset markets arose to offer current claims on future receipts from helicopter flights, individuals would be able to incorporate titles to random future money issues directly into their initial portfolios Such money "futures" contracts would have the same effect on consumers' decisions as would period 1 money balances under distribution rule *a*, both assets bring future money mintings under the consumer's ownership

Consideration of Realism

The helicopter mint is a useful pedagogic device, but it bears little resemblance to the actual method of money issuance currently in vogue In fact, the present issuer uses a sizable fraction of new fiat money to buy helicopters and pilots (and other real goods and services) with the purpose of dispensing objects more lethal than paper bills The acquisition of real commodities through fiat money issuance complicates life for the economist as well as for the citizen because it necessitates the consideration of commodity production A realistic money supply rule just cannot be appended to the individual's first-order demand conditions Rather, it must encompass the rules used by government to select commodities for purchase with newly minted money and the rules used to determine how those commodities will be reallocated among individuals

In the simple world where fiat money was distributed directly to individuals, a dynamic equilibrium between the rate of monetary expansion and the rate of inflation was found when individuals received money or were relieved of money in proportion to their just previous nominal money balances This result was brought about by rational and accurate forecasters who discounted in advance the future production of new money and caused current balances to fluctuate in value with changes in their explicit claims to future money issues The fact of government dispensing new money indirectly, by buying and distributing commodities, does not logically require a change in the presumption of rationality and in the forecasting ability of individuals [15] Thus, if government-purchased commodities are redistributed to individuals in proportion to their former nominal money balances, and if those commodities are marketable (transferable), advanced discounting of future receipts of goods would increase the current demand for nominal money balances For example, a situation wherein commodities purchased by the government in each period are distributed proportionately to money balances held at the beginning of the first period is effectively the same as helicopter rule *a* and will bring the same inflation sequence, namely, a very large inflation in the first period and zero inflation thereafter

Again speaking of things as they actually are, the government neither distributes new money nor taxes old balances proportionately It may indeed distribute goods and services roughly in relation to wealth, but wealth is composed of many other assets besides money, and for an exact simultaneous relation to exist between money and inflation the link of new money issues to old balances must be exactly proportional

Little additional time need be devoted to the possibility of private asset supply Since private supplies are equivalent in effect to short sales, they have been implicitly accommodated on the demand side and thus will alter no conclusions by being explicitly considered If legal restrictions on privately validated, non-interest-bearing currency were to be removed, their issuers would assume the status of this paper's government, and the time sequences of that currency's commodity value would be determined by the identical mechanism Interest rates contracted in that currency would be governed similarly

Finally, we must hedge the present analysis against possible real effects of money If a position of underemployment (short-run) equilibrium can be alleviated by a monetary expansion, the link between nominal money supply and commodity price level becomes much more tenuous To put it simply, in a condition of unemployment, newly minted money could be used to coax production from idle capital and labor, at least this seems to be the prevailing opinion among macroeconomists If this is indeed true, then no inflation tax would inure to money holders, and the resulting rates of inflation would be lower than the rates implied earlier in this paper regardless of the money distribution rule

---

[15] However, it most definitely changes the achievable level of utility of each individual since it alters the productive mix of commodities by favoring those the government happens to choose for its reallocation endeavors Naturally, I have neglected many important matters here such as the production of some goods, say nuclear weapons, which because of difficulty in charging benefits, external effects, nontransferability, etc, would not have existed without the government's demand for them

References

Fisher, Irving *The Theory of Interest* New York Macmillan, 1930

Friedman, Milton "The Optimum Quantity of Money" In *The Optimum Quantity of Money and Other Essays* Chicago Aldine, 1969

Kushner, Harold J "On Stochastic Extremum Problems Calculus" *J Math Analysis and Appl* 10 (April 1965) 354–67

Long, John B, Jr "Consumption Investment Decisions and Equilibrium in the Securities Market" In *Studies in the Theory of Capital Markets*, edited by Michael C Jensen New York Praeger, 1971

Meltzer, Allan H "Discussion" *J Finance* 25 (May 1970) 450–53

The Honest Government's Guide to the Revenue from the Creation of Money

Leonardo Auernheimer

Texas A & M University

The analysis of the revenue from the creation of fiat money is usually carried out in terms of comparative statics. Given the value of relevant parameters, the rate of inflation which maximizes revenue is characterized as the one yielding the highest stream of revenue once all adjustments have been completed: either some rate of monetary expansion is set, and the results corresponding to the long-run values of the variables are analyzed (Friedman 1971), or the inflation rate is fixed at the outset [illegible] adjustments taken to be instantaneous, but assuming away any effect of the once-and-for-all changes in real cash balances on the variable to be maximized (revenues), that is, assuming that those changes are met by once-and-for-all changes in the price level (Bailey 1956). Of course, both procedures amount to the same thing.

This is, indeed, the correct analysis in processes in which the transition period can be assumed away because the long-run value of the variable to be maximized is reached asymptotically from the very beginning, so that the highest stream at any point is also the stream with the highest present value at the initial point. This is not so in the case of the revenue from the creation of money.

The purpose of this paper is to show, first, how explicit consideration of revenues accruing during the adjustment period (or "at" the adjustment, when it occurs instantaneously) is essential for the understanding of the essence of money as a capital good, and second, how for a particular adjustment behavior—instantaneous adjustment—taking due account of once-and-for-all changes produces results very different from those reported in the literature. In general, lower "revenue-maximizing" rates of inflation are called for, and in particular, those rates are indepen-

---

This paper was written at the University of Chicago, and it is based on a chapter of my dissertation. I am thankful to S. Fischer, R. Gordon, C. A. Rodriguez, L. Sjaastad, and R. Zecher, and in particular to M. Friedman, who asked the right question. A pitiless comment by H. G. Johnson and a set of careful suggestions by an anonymous referee forced me to restate the argument in what I hope is a clearer fashion. After this paper was submitted for publication, an unpublished paper by Charles D. Cathcart was brought to my attention, in which some of the points made here are independently developed by Cathcart.

dent of the rate of growth of the economy. For expository purposes, we first proceed to demonstrate the second point

I. The Assumptions

The explicit consideration of transitional effects makes it necessary to specify the assumptions very precisely For the particular case we examine, they are (*a*) When some rate of inflation (π) is announced, it is believed by the public and "implemented" immediately by the government (*b*) The adjustment of actual to desired cash balances is so quick that it is proper to treat it as being instantaneous (*c*) No once-and-for-all changes in the price level are allowed, that is, when a "lower than the preexisting" rate of inflation is announced, the government steps into the market and "sells" the additional real balances desired by the public Likewise, when the rate of inflation is increased, it is prepared to "repurchase" the excess of existing over desired real balances. This means that no once-and-for-all "subsidies" or "capital levies" are imposed upon the public

Assumptions (*a*) and (*b*) are the same as the ones used, for example, by Bailey (1956) Assumption (*c*) is critical, and requires some elaboration

Assumption (*c*) does not present any difficulty for the case where some lower (than the preexisting) rate of inflation is announced, and the government reaps the once-and-for-all increase in total wealth, there is no reason why it should not do so, interested as it is in maximizing revenue The question is, Why should government proceed in a symmetric fashion when a higher rate of inflation is announced, that is, to intervene in the market and sell goods so that no once-and-for-all change in the price level occurs?

There are at least two reasons for assuming so In the first place, a "once-and-for-all" change in the price level is a short-hand expression for a very short period of high inflation, strictly, there is some "cheating" by part of the government in allowing that to happen Here is where the adjective "honest" applies The same assumption is used by Foley and Sidrausky (1971) in analyzing the effect of changes in the rate of inflation, in the context of their model [1] The second reason is that, if the symmetry is not imposed, the model becomes highly unstable when the other two assumptions are maintained The reasons for this are explained in Section IV below

---

[1] "The change in the target rate of inflation is reflected by a change in the slope of the p_m (semilogarithmic) line at time T, but there is no jump in p_m at time T If, when the government announces a change in the rate of inflation, it does not want to disappoint private expectations about the course that p_m will follow, then no jump in p_m can be allowed at $t = T$ Otherwise, the rate of change of p_m at $t = T$ would be infinity" (Foley and Sidrausky 1971, p 190)

Finally, it is worthwhile to mention that at least assumption (*c*) acquires full empirical relevance in cases where the preexisting rate of inflation is higher than the "optimum" rate, as described below

II. The Analysis

In what follows, for simplicity, I will use a Cagan-type function of the demand for money and assume constant income elasticity (not necessarily equal to unity) I also assume, as it is usually done, that government is the only issuer of money

Total real revenues per (infinitely small) period are defined as

$$R_t = \frac{dM}{dt}\frac{1}{P} = \frac{dM}{dt}\frac{1}{M}\frac{M}{P} = \rho m_t, \tag{1}$$

where ρ is the (proportional) rate of monetary expansion, and m is real cash balances

By differentiating the identity $m = M/P$,

$$R_t = \rho m_t = m_t\pi_t + \dot{m}_t,$$

where π_t is the rate of inflation and a dot over the variable denotes the time derivative

If the desired stock of money—identically equal to the actual stock is

$$m_t = A\ e^{\gamma t}\ e^{-b(r+\pi)}, \tag{2}$$

where γ = a "growth factor" (equal to $n + g\eta$, for n = rate of population growth, g = rate of growth of per capita income, η = income elasticity of demand), r = "the" real interest rate, A, $b > 0$, parameters then $\dot{m}_t = \gamma m_t - b m_t \dot{\pi}$ If adjustments are instantaneous, $\dot{\pi} = 0$ after $t = 0$, so that the flow of revenue after $t = 0$ becomes $R_t = m_t\pi + \dot{m}_t = m_t(\pi + \gamma)$ Assuming that the future is discounted by government at the prevailing rate of interest r, at any time $t = 0$ the present value of the revenue from the creation of money, for some constant rate of inflation, is given by

$$V_{t=0} = \int_{t=0}^{\infty} m_t(\pi + \gamma)e^{-rt}\,dt + (m_{t=0+} - m_{t=0}), \tag{3}$$

where $m_{t=0}$ are cash balances at the preexisting rate of inflation, π_0 and $m_{t=0+}$, cash balances corresponding to the new rate of inflation, π, that is, just after the change

The first term in expression (3) is the present value of the flow of revenue obtained at each period after $t = 0$, this is the term usually considered in the literature The second part is the once-and-for-all "sale" (positive or negative, according to whether $m_{t=0+} \gtrless m_{t=0}$, or $\pi \lesseqgtr \pi_0$) of real cash balances which takes place at $t = 0$

Given the assumed form of the demand for money, equation (3) becomes

$$V_{t=0} = Ae^{-b(r+\pi)}(\gamma + \pi) \int_{t=0}^{\infty} e^{(\gamma - r)t}\, dt + Ae^{-br}\left(e^{-b\pi} - e^{-b\pi_0}\right) \tag{4}$$

Assuming $r > \gamma$, so that the integral converges to some finite value,[2]

$$= \frac{Ae^{-b(r+\pi)}(\gamma + \pi)}{(r - \gamma)} + Ae^{-br}\left(e^{-b\pi} - e^{-b\pi_0}\right) \tag{5}$$

Differentiating with respect to the rate of inflation,

$$\frac{\partial V_{t=0}}{\partial \pi} = Ae^{-br}\ e^{-b\pi}\left(-\frac{b}{\gamma - r} + \frac{1}{\gamma - r} - \frac{b\pi}{\gamma - r} + b\right),$$

and equating to zero,

$$1 - b\gamma - b\pi + b(\gamma - r) = 0,$$

$$\pi = \frac{1}{b} - r \tag{6}$$

That is, the revenue-maximizing rate is independent of the rate of growth of the economy[3]

From expression (6) it follows that, to the extent that the basic assumptions are met, the usual analysis overestimates the "revenue-maximizing" rate (and, consequently, errs also in the calculation of magnitudes such as maximum revenues, ratio of maximum revenues to total income, etc) The interested reader can easily calculate the extent of the overestimation for particular values of the parameters

The overestimation is larger the lower the value of the growth factor, γ This result is consistent with a particular interpretation of the approach taken in this paper In the usual analysis, the concern is with the revenue derived from keeping the new desired real balances constant, and from providing future "generations" with the same balances,[4] but not from taking charge of the difference in the real stock desired by existing people at $t = 0$. In other words, we consider existing people just like future "generations."[5] The larger the growth factor, γ, the smaller the weight of instantaneous "sales" in "total new sales" as a source of revenue—*as opposed to the revenue derived from a tax on existing balances*.

---

[2] The case when $r < \gamma$ confronts us with the same situation where an individual owns a stream of income which grows at a constant rate, larger than the rate of interest of the economy Such an individual is infinitely wealthy, provided he can borrow ad infinitum If he tries to consume ad infinitum, eventually the rate of interest will rise The relationship between the rate of growth and the rate of interest in the steady state has been ignored here

[3] An alternative proof, along the lines of the calculus of variations, is immediate

[4] Unless income per capita does not grow, so that the "growth factor" γ reflects only population growth, the term "generation" has a special meaning in this particular interpretation It refers to some combination of "new bodies" and "new purchasing power "

From expression (5), a geometric interpretation is straightforward. At any instant $t = 0$, the difference between the present value of future total revenue (at the preexisting rate π_0) and the present value of the revenue from some alternative rate, π, can be expressed as

$$(\Delta\Gamma)_{t=0} = (m_{t=0+} - m_{t=0}) + \frac{m_{t=0+}(\gamma + \pi)}{r - \gamma} - \frac{m_{t=0}(\gamma + \pi_0)}{r - \gamma} \tag{7}$$

The first term in expression (7) is the once-and-for-all gain (loss) from changing to a lower (higher) rate of inflation, the second and third, the present values of the two alternative streams. This can be reduced, after some transformations, to

$$\begin{aligned}(\Delta\Gamma)_{t=0} &= \left(\frac{1}{r - \gamma}\right)[m_{t=0+}(r + \pi) - m_{t=0}(r + \pi_0)],\\ &= \left(\frac{1}{r - \gamma}\right)[(r + \pi)(\Delta m) + (\Delta\pi)(m_{t=0})],\end{aligned} \tag{8}$$

where $(r + \pi)(\Delta m) = (\text{area } A) + (\text{area } B)$ in figure 1, and $(m_{t=0})(\Delta\pi) = (\text{area } C)$. The increment in the total present value will be zero when area $(A + B)$ and area C are equal, independently of the value of γ. The analysis can be carried out at any point in time, that is, for any position of the demand curve $I - L$.

Notice that this can also be expressed as a flow, by multiplying the total marginal present value (8) by the rate of interest; in particular, when $\gamma = 0$ it becomes

$$r(\Delta\Gamma)_{t=0} = (r + \pi)(\Delta m) + m_{t=0}(\Delta\pi).$$

Observe also that these results are the same as what would be obtained when revenue per period is taken to be equal to $m(r + \pi)$; the meaning of this will become clear in Section III.

III A Digression on the Economics of the Supply of Capital Goods

In a recent paper (Friedman 1971), Friedman asks the question

---

[5] I am indebted to Stanley Fischer for this interpretation.

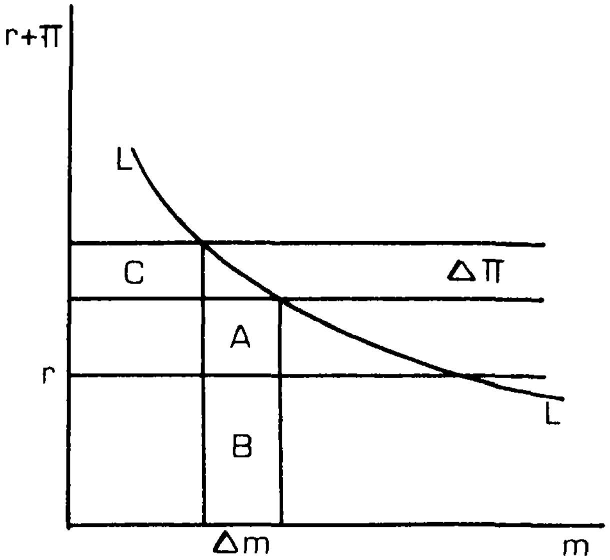


Fig. 1

> It has always seemed natural to assimilate a government monopoly of fiat money issue with a private monopoly of a product produced at zero cost: Cournot's mineral spring. Yet we have just demonstrated that this is not correct. The owner of the spring will maximize his proceeds by charging a price at which the price elasticity of demand is unity. And this will be true whether the economy is stationary or growing. Why is this not also true for the government which can issue money at zero marginal cost? Why will it not also maximize its revenue by pricing at the unit-elasticity point on its demand curve, regardless of whether there is growth or not?
>
> The answer is because there are two different prices that are relevant to money issue—one, the goods and services that are given up to acquire a dollar, the other, the number of cents per dollar per year that *the money holder must spend to keep his* real balances constant. Only the latter is analogous to the price of the mineral water. [Friedman 1971, pp. 854–55]

Then, he suggests an analogy with the "monopolistic producer and servicer of . . . durable equipment who sells the equipment but charges an annual fee for servicing it."

Our analysis points toward a somewhat different interpretation. The correct analogy seems to be with Friedman's *producer and servicer* of durable equipment, provided that (1) the equipment has a *fixed sale* price (equal to unity, in our money case), (2) the producer is committed to "repurchase" the equipment at that fixed price. An alternative example, amounting to the same thing, is the case of a monopolist renting a capital good (say, telephones), at an explicit rental per unit per year (π), and charging a fixed and refundable deposit. At any period, total revenues are made up of the explicit rental on the existing stock ($m\pi$ in our money case), and of the net receipts of deposit funds ($\dot{m}$ in our case). The analogy in this case is exact, and we need only to introduce assumption (*c*) (Section I above) when adjustments are instantaneous. In both cases (telephones and money), what is bought by the public is services, and the total price paid for those services is made up of the "explicit rental," π, plus the "implicit rental," the real interest rate.[6] The correct maximization procedure, then, is with respect to the total paid for the services sold; this yields a result identical to that obtained from the production and sale of any other product, that is, independent of the rate of growth of the economy.[7]

IV. The Consequences of Relaxing Some Assumptions

The analysis of Section II is limited by the extreme restrictiveness of assumptions (*a*)–(*c*), discussed in Section II. We proceed now to show the extent to which the relaxation of each of them would modify the results.

1. If assumptions (*a*) and (*b*) are maintained, and assumption (*c*) is not imposed, the model is unstable. Consider a situation where, for each discrete change in the rate of inflation, the government is allowed to bear (reap) or not the once-and-for-all losess (gains). If so, the optimal behavior will be to reap the gains when π is lowered, and let the public bear the losses when π is increased. Take, as an initial position $\pi = (1/b) - r$ (assuming no growth, for simplicity); now, it will be profitable to switch to the higher rate of inflation $\pi = 1/b$, to which a higher stream is attached, since there will be no once-and-for-all loss to (more than) compensate for the gain in the flow (i.e., only the first term in expression [3] is relevant). If, instead, $\pi = 1/b$ is taken as the initial position, then it is profitable to change to a lower rate, $\pi = (1/b) - r$, as the once-and-for-all gain more than compensates for the loss in the

---

[6] The theory of the optimum quantity of money makes it clear that this last component is a cost for money holders; the literature on the revenue from the creation of money does not, in general, see it as a "revenue" for the money issuer. For an exception, see Johnson (1969*a*, 1969*b*).

[7] The question is somewhat similar to the one arising in the case of a monopolist selling some capital good, where the (exponential) rate of depreciation can be controlled by the producer. In a recent paper (1970), Swan has shown that some erroneous conclusions in the literature were derived precisely because of ignoring once-and-for-all changes in the existing stock—changes which, as a purely mathematical matter, cannot be captured by static maximization procedures when capital goods, and their price—instead of services and rentals—are considered.

present value of the future stream (i e , now, both terms in expression [3] are relevant) The result is that there will always be incentives for "jumping" from any of the two rates of inflation to the other Obviously, we cannot expect the other two assumptions to survive under these circumstances

2 Assumptions (*b*) and (*c*) are twin brothers if a lag in the adjustment of actual to desired cash balances is allowed for—removal of assumption (*b*)—then assumption (*c*) is no longer necessary, and the analogy with the private monopolist as described in the previous section is exact It can be easily shown, indeed, that the "instantaneous-adjustment" case is but a particular limiting instance of the more general case

3 Removal of assumption (*a*) presents different difficulties The analogy with the private monopolist as described above is preserved in its formal presentation, but no longer can the "control variable" be taken as being the same while our monopolist can set the explicit rental at every period, the same is not true of the fiat money issuer But, in any case, statements about the optimum (constant) value of the explicit rental which are valid for any of the two remain valid for the other

4 If no restrictive assumptions are imposed, then no rule on optimum behavior can be deduced a priori (even if, for example, the rate of monetary expansion is forced to be constant from the very beginning) without knowing the exact form of the adjustment and the expectations equations Of course, it will be true that for the whole adjustment period total revenues are

$$\int_{t_0}^{t_1} R_t \, dt = \int_{t_0}^{t_1} m_t \pi_t \, dt + \int_{t_0}^{t_1} m_t \, dt,$$

where $t_1 - t_0$ is the duration of the adjustment period For any change in the equilibrium rate of inflation, from the demand for money function we know precisely the sign and magnitude of the second integral The problem, though, is that nothing can be said regarding the magnitude of the first integral Furthermore, the problem is complicated when earlier flows carry a larger weight than later ones (in particular, if the problem is analyzed in the context of a model allowing for a lag in the rate of inflation after a stepwise change in the rate of monetary expansion)

V. Conclusions, and Two Implications

We have shown how explicit consideration of changes occurring at the transition allows, in general, for a better understanding of some characteristics common to both money and other capital goods For a particular adjustment process (instantaneous adjustment), we have deduced a "revenue-maximizing" rule different from the one derived in the literature

Even though the main point in this note is theoretical, at least two empirical considerations follow. First, that there may be substantial once-and-for-all gains (for both growing and stationary economies) from a reduction in the rate of inflation, gains which may accrue almost immediately, provided that adjustments are fast and government can effectively persuade the public about its intentions, these once-and-for-all gains may be large enough to compensate for a fall in the subsequent permanent stream. This is still more relevant in cases where some time is needed either for decreasing the level of public expenditure or for raising revenues from conventional taxes.

Second, the argument shows how, in principle, the provision of the optimum quantity of money is "self-financed" (in the sense that there is no need to impose conventional taxes for the payment of interest on cash balances) if at every instant (1) the first component of total revenues ($\dot{m}$) is used to purchase real capital and integrate an "investment fund" (presumably in the form of equity) and the second component ($m\pi$) is used to pay interest on real balances, (2) the interest received on the "investment fund" (mr) is used to complete interest payments in real [illegible]h balances. Of course, if in the past government has consumed part or all of the "investment fund," and at some point it decides to start paying interest on money, then it would need "outside financing." But if this is the case and the economy is growing, the proportion of interest paid on money which needs to be "outside financed" decreases over time, and eventually becomes negligible after a few years.[8]

References

Auernheimer, Leonardo. "Essays in the Theory of Inflation." Ph.D. dissertation, Univ. Chicago, 1973.

Bailey, Martin J. "The Welfare Cost of Inflationary Finance." *J.P.E.* 64 (April 1956): 93–110.

Foley, D. K., and Sidrauski, M. *Monetary and Fiscal Policy in a Growing Economy.* New York: Macmillan, 1971.

Friedman, Milton. "Government Revenue from Inflation." *J.P.E.* 79 (July/August 1971): 846–56.

Johnson, Harry G. "Inside Money, Outside Money, Income, Wealth, and Welfare in Monetary Theory." *J. Money, Credit, and Banking* 1, no. 1 (February 1969): 30–45. (*a*)

———. "A Note on Seigniorage and the Social Saving from Substituting Credit for Commodity Money." In *Monetary Problems of the International Economy,* edited by R. A. Mundell and A. Swoboda. Chicago: Univ. Chicago Press, 1969. (*b*)

Swan, Peter L. "Durability of Consumption Goods." *A.E.R.* 60 (December 1970): 884–94.

---

[8] For further elaboration of this point, and some hypothetical calculations, see Auernheimer (1973, chap. 1).

Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utilities: A Note

Antonio Camacho

University of Illinois, Chicago Circle

Jon Sonstelie

Northwestern University

In 1955 John Harsanyi proposed three appealing postulates for social choice under uncertainty and tried to show (Theorem V, p. 314) that these postulates lead to a social welfare function which is a weighted sum of the utility functions of the individuals.

The importance of this result is that it shows that a rather weak and plausible ethical postulate (Postulate C below) suffices to establish a cardinal social welfare function which is a weighted sum of the individuals' cardinal utility functions.

However, as presented by Harsanyi, his conclusion can be applied only to a restricted class of cases which we do not consider broad enough *to incorporate many important situations of social choice that may* appear in the real world.

In this paper we will attempt to improve Harsanyi's result in two directions. First, instead of his Postulate C we will use a new Postulate C′, which is even weaker than Postulate C. Second, our proof of the existence of a cardinal social welfare function that is a weighted sum of the individual utility functions will *not depend on certain restrictive* assumptions utilized in Harsanyi's proof, and consequently our result can be applied to a broader spectrum of situations. Finally, we want to note that the mathematics employed in our proof is very simple indeed.

We now present Harsanyi's postulates and Theorem V. Then, Postulate C′ will be introduced and our stronger theorem (Theorem V′) proved.

Harsanyi's three postulates are:

Postulate A. Social preferences R satisfy Marschak's (1950) Postulates I, II, III′, and IV.

Postulate B. Individual preferences $R_i (i = 1, \ldots, n)$ satisfy the same four postulates.

Postulate C.—If two prospects P and Q are indifferent from the standpoint of every individual, they are also indifferent from a social standpoint.

Let $X_1,\quad, X_j,\quad, X_m$ represent the m sure prospects available to society From Postulate A it follows that there exists a social welfare function W, representing the social preferences, that satisfies the following property (α) If the value of W corresponding to the sure prospect $X_j (j = 1,\quad, m)$ is w_j, then the value of W for any prospect $P = (p_1,\quad, p_j,\quad, p_m), p_j \geq 0,$

$$\sum_{j=1}^{m} p_j = 1,$$

is

$$W(p_1,\quad, p_j,\quad p_m) = \sum_{j=1}^{m} p_j w_j$$

This social welfare function is unique up to a positive linear transformation Similarly, from Postulate B it follows that for each individual $i (i = 1,\quad, n)$ there exists a utility function U_i, representing his preferences R_i, and satisfying property (α)

$$U_i(p_1,\quad, p_j,\quad, p_m) = \sum_{j=1}^{m} p_j u_{ij},$$

where u_{ij} is the value of U_i that corresponds to the sure prospect X_j This utility function is unique up to a positive linear transformation

Harsanyi's Theorem V can now be stated as follows If the individual preferences $R_i (i = 1,\quad, n)$ satisfy Postulate B, and the social preferences R satisfy postulates A and C, and the corresponding utility functions U_i and the social welfare function W are chosen in such a way that for a given prospect, all take the value zero, then the social welfare function W can be represented as a weighted sum of the individual utilities of the form

$$W = \sum_{i=1}^{n} a_i U_i,$$

where the a_i's are constant real numbers, that is, they do not depend on the prospects P

A close look at Harsanyi's paper suffices to show that his proof of Theorem V relies on the existence of prospects such that, after choosing a common origin for the individual utility functions, the utility of each of these prospects is one for a given individual and zero for the rest of the individuals in the society For those situations where there are prospects S_k $(k = 1,\quad, n)$, such that

$$U_i(S_k) = \begin{cases} 1, & \text{if } i = k, \\ 0, & \text{if } i \neq k, \end{cases} \qquad (\beta)$$

Harsanyi's proof and his interpretation of the coefficients a_i's are both correct But in those case where, after choosing a common origin for all the individual utility functions, there is no prospect whose utility is one for a given individual and zero for the others, neither his proof nor his interpretation of the coefficients a_i applies

We believe that in those situations where the possible social states that the individuals face are probability distributions of alternative income distributions, condition (β) is satisfied But for other important social choice situations that are likely to appear in the real world, such as public education and other public projects, we may very well have the case where the preferences of the individuals are such that condition (β) does not hold The simplest of these cases is, of course, that in which the preferences of all individuals with regard to the alternative social prospects are the same But other less trivial situations may also appear

The proof that we will present below is not based on the existence of prospects with utility equal to one for an individual and zero for the others and makes use only of a Postulate C′ that is weaker than Harsanyi's Postulate C Let $\bar{P}$ be any given prospect with all its components strictly positive $\bar{P} = (\bar{p}_1, \quad , \bar{p}_j, \quad , \bar{p}_m)$, $\bar{p}_j > 0, j = 1, \quad , m$,

$$\sum_{j=1}^{m} \bar{p}_j = 1$$

Postulate C′—If a prospect P is indifferent to $\bar{P}$ from the standpoint of every individual, then it is also indifferent to $\bar{P}$ from a social standpoint

We will now prove the following Theorem V′ If the individual preferences R_i ($i = 1, \quad , n$) satisfy Postulate B, and the social preferences R satisfy postulates A and C′, and the corresponding utility functions U_i and the social welfare function W are chosen in such a way that for a given prospect all take the value zero, then the social welfare function W can be represented as a weighted sum of the individual utilities of the form

$$W = \sum_{i=1}^{n} a_i U_i,$$

where the a_i's are constant real numbers, that is, they do not depend on the prospects P

Proof—Observe first that in view of property (α), to prove that

$$W = \sum_{i=1}^{n} a_i U_i,$$

it suffices to show that

$$w_j = \sum_{i=1}^{n} a_i U_{ij},$$

for $j = 1, \quad , m$

Observe also that if the theorem holds when a given prospect is chosen as the common origin of the utility functions and the social welfare function, then it also holds if any other prospect is chosen as the common origin. Thus, without loss of generality, let the prospect $\bar{P}$ be chosen as the common origin. Consider now the two following systems of equations:

$$\begin{cases} u_{11}p_1 + \ldots + u_{1m}p_m = 0 \\ \ldots \\ u_{n1}p_1 + \ldots + u_{nm}p_n = 0, \end{cases} \tag{1}$$

$$\begin{cases} w_1p_1 + \ldots + w_mp_m = 0 \\ u_{11}p_1 + \ldots + u_{1m}p_m = 0 \\ \ldots \\ u_{n1}p_1 + \ldots + u_{nm}p_m = 0 \end{cases} \tag{2}$$

Clearly we have the following:

i) Prospect $\bar{P} = (\bar{p}_1, \ldots, \bar{p}_j, \ldots, \bar{p}_m)$ is a solution of both system (1) and system (2), and $\bar{p}_j > 0$, for $j = 1, \ldots, m$. Thus both systems have a common solution with all its components strictly greater than zero. The set of all solutions of system (1) which are prospects is precisely the set of all prospects which are indifferent to prospect $\bar{P}$ from the standpoint of every individual. The set of all solutions of system (2) which are prospects is precisely the set of all prospects which are indifferent to $\bar{P}$ from the standpoint of every individual and from the standpoint of society.

ii) Any solution of system (2) is also a solution of system (1), since system (2) is obtained from (1) by adding the equation $w_1p_1 + \ldots + w_mp_m = 0$.

iii) If $(p_1, \ldots, p_j, \ldots, p_m)$, not necessarily a prospect, is a solution of (1) and $p_j \geq 0$ for $j = 1, \ldots, m$, then

$$\left(\frac{p_1}{\sum_{j=1}^{m} p_j}, \ldots, \frac{p_m}{\sum_{j=1}^{m} p_j}\right),$$

which is a prospect, is a solution of (1) and by Postulate C′, it must also be a solution of (2), which implies that $(p_1, \ldots, p_m)$ is a solution of (2).

From (i), (ii), and (iii) it follows that (*a*) $(\bar{p}_1, \ldots, \bar{p}_m)$, where $\bar{p}_j > 0$ for all j, is a solution of both system (1) and system (2), and (*b*) any $(p_1, \ldots, p_m)$, with $p_j > 0$ for all j, which is a solution of (1), is also a solution of (2), and vice versa. But (*a*) and (*b*) imply that systems (1) and (2) are equivalent, that is, every solution of (1) is a solution of (2) and vice versa.[1] Since system (2) is derived from system (1) by adding the equation $w_1 p_1 + \cdots + w_m p_m = 0$, it follows that

---

[1] Let $(q_1, \ldots, q_j, \ldots, q_m)$ be any solution of (1). We can write $q_j = \bar{p}_j + \lambda\varepsilon_j$, where $\bar{p}_j + \varepsilon_j > 0$ for $j = 1, \ldots, m$. Then, since $(\bar{p}_1, \ldots, \bar{p}_m)$ is a solution of (1), $(\lambda\varepsilon_1, \ldots, \lambda\varepsilon_m)$, $(\varepsilon_1, \ldots, \varepsilon_m)$, and $(\bar{p}_1 + \varepsilon_1, \ldots, \bar{p}_m + \varepsilon_m)$ are also solutions of (1). Since $\bar{p}_j + \varepsilon_j > 0$ for all j, it follows that $(\bar{p}_1 + \varepsilon_1, \ldots, \bar{p}_m + \varepsilon_m)$ is a solution of (2). And since $(\bar{p}_1, \ldots, \bar{p}_m)$ is a solution of (2), it follows that $(\varepsilon_1, \ldots, \varepsilon_m)$, $(\lambda\varepsilon_1, \ldots, \lambda\varepsilon_m)$, and $(\bar{p}_1 + \lambda\varepsilon_1, \ldots, \bar{p}_m + \lambda\varepsilon_m) = (q_1, \ldots, q_m)$ are also solutions of (2). In the same way we can show that any solution of (2) is a solution of (1).

$$w_j = \sum_{i=1}^{n} a_i u_{ij},$$

as desired.

References

Harsanyi, John. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *J.P.E.* 63 (August 1955): 309–21.

Marschak, Jacob. "Rational Behavior, Uncertain Prospects, and Measurable Utility." *Econometrica* 18 (April 1950): 111–41.

On Killing off the Market for Used Textbooks and the Relationship between Markets for New and Secondhand Goods

H Laurence Miller, Jr

University of Hawaii

Are secondhand copies of textbooks a drain on publishers' and authors' profits? Publishers' representatives have no doubts on the matter I have talked with several about the "problem" and about some companies' efforts to reduce secondhand sales One is said to have aimed for "instant destruction" by including answer pages in books to be torn out and handed in to the instructor, decreasing the value of used copies That the publishers' representatives have some support from professional economists is clear from this statement in the preface to a widely used textbook in price theory "Since everyone knows the basic reason for a revised edition is to kill off the existing used book market, it would be idle to suggest otherwise "[1]

It is sometimes said, on the other hand, that the initial price captures the present value of all subsequent transactions so that used items do not in fact compete with new ones The buyer of a newly produced diamond pays a price consistent with what the diamond can be sold for to others including members of later generations

Textbooks and diamonds do differ The cost of producing additional textbooks after setup costs falls more dramatically than the cost of producing additional diamonds It costs more to produce a hardcover book than a paperback, whereas the cost of producing a diamond is independent of its life and number of uses And a used textbook, unlike a used diamond, is seldom a perfect substitute for a new one But are these differences important?

Would publishers and authors be better off if they could kill off the secondhand market for their products? The problem is more complex than some appear to suppose as well as analytically similar to problems arising in other contexts in economic theory [2]

I am indebted to Armen Alchian, Milton Friedman, and Akira Takayama for discussions of an earlier version of this paper They bear no responsibility for the result

[1] Ferguson 1969, p vii

[2] See especially the problem "Alcoa The Secondhand Market and Market Position," in Friedman (1962, pp 278–79) I am indebted to Friedman for pointing out the similarity Some matters discussed in this paper parallel those in the peak pricing literature (cf also Hirshleifer 1971)

I

Consider an idealized situation (1) Transaction costs are zero (2) A good has a technically given service life equal to two possible uses or time periods (3) Costs of production per unit are constant (4) All goods produced new, X_1, survive into the second period to provide potential second period supply, X_2, so that under certain conditions secondhand sales in any period would equal the previous period's sales of new units If optimal sales in the second period are less than first-period production, storage costs are prohibitive, disposal costs negligible (5) New goods are preferred to used goods at identical prices Though some of these assumptions conflict with the facts relating to production of books and other items, they are initially harmless and can be modified

Let some buyers purchase new units, keep them for one period, then trade them in on new ones, others buy secondhand units, keep them for the second time period, discard them, and buy another set of traded-in units Alternatively, one group rents the good for the first part of its life, the other rents it for the second One can think of an automobile that lasts exactly 2 years and consumers in the market year after year for the services rendered by automobiles One cannot think realistically of individuals in the market year after year for the new edition of a textbook Although a new set of individuals enters the market for textbooks every quarter, semester, or year (and some students do fail), the previous period's demanders leave the market But this difference between the market for textbooks and some other goods is not crucial

Assume, finally, that the demand for new items is $P_1 = P_1(X_1)$ and demand for used items is $P_2 = P_2(X_2)$ so that the demands are independent The cross-partial elasticities of demand $(\partial X_j/\partial P_i) = 0$, $i \neq j$ These demands appear as D_1 and D_2 in figure 1 This simple version of our problem is illuminating though too restricted to fully answer the questions asked above

Figure 1 portrays long-run equilibrium in a competitive industry [3] (We can ignore discounting and the problems constant costs raise regarding competitive supply) The *vertical* summation of D_1 and D_2 is D_T The equilibrium rate of production and new sales, X_1^*, equals secondhand sales, X_2^*, P_1^* and P_2^* would be the effective annual prices paid by owner-consumers or renters, $P_1^* + P_2^* = C$ First owners pay a gross price C and receive a trade-in allowance of P_2^* one period later, $C - P_2^* = P_1^*$, $P_2^* - 0 = P_2^*$ Transaction costs and scrap value are zero Vertical summation is a familiar step in analyzing problems in joint production—here the services the book renders in turn to first and second users

Figure 2 applies this argument to a monopolist who maximizes profits by equating marginal cost $C'(X_1)$—which equals average cost, $C(X_1)/X_1$

---

[3] Figure 1 is identical with one in Miller (1960, p 251) That analysis is correct so far as it goes I simply adopted Fox's assumption of independent markets

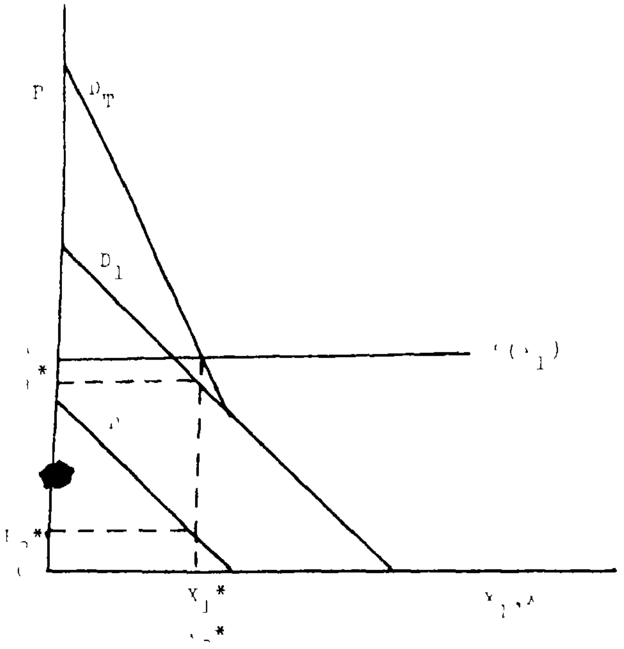


Fig. 1

and the vertical sum of the marginal revenue functions at output and sales $X_1^* = X_2^*$. First owners pay a gross price equal to $P_1^* + P_2^*$, a net price of P_1^*. Second owners pay a net price P_2^*. Alternatively, P_1^* and P_2^* are the full rental prices to renter-users. Here $P_1^* + P_2^* > C$.

With the cost-revenue situation in figure 2, X_1^* can equal X_2^* with marginal revenue in both markets nonnegative. Equal sales in both markets could lead to a negative marginal revenue in the secondhand market. No monopolist would choose the output X^c in figure 3 where $MR_2 < 0$ unless forced to do so by a regulatory agency. Used units would be undesired by-products rather than desirable joint products. The rule is that marginal cost must equal total marginal revenue, $MR_1 > 0$, $MR_2 \geq 0$, with $MR_2 = 0$ replacing $MR_2 < 0$ for secondhand sales exceeding X_2^*. Optimum sales will be X_1^* and X_2^*, $X_1^* > X_2^*$. The quantity $X_1^* - X_2^*$ will be stored or disposed of depending upon relevant costs and revenues.

These remarks apply to a monopolist who rents his product. He need not rent unprofitable secondhand units. If he sells units to owner-consumers who will have X_1^* units to sell in secondhand markets, it will pay him to buy back $X_1^* - X_2^*$ secondhand units at a price P_2^*. First owners will anticipate this policy and pay an initial price that counts on a trade-in value of P_2^* for all units.

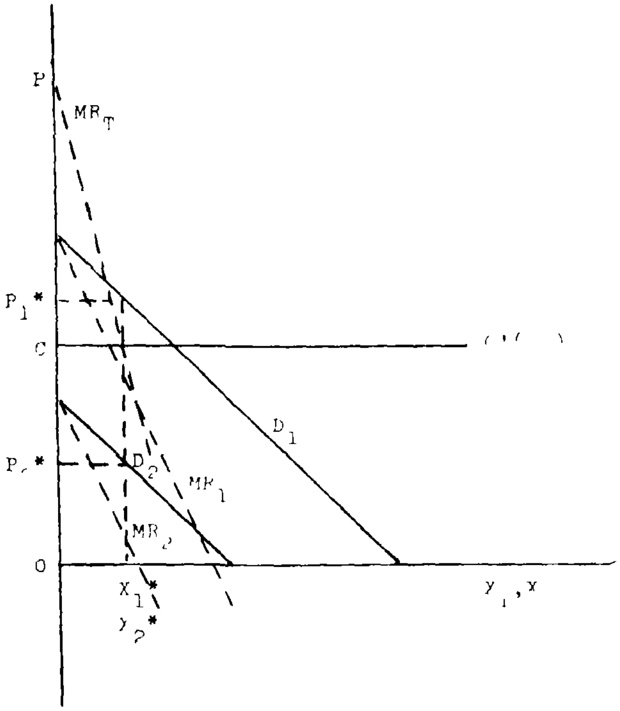


FIG. 2

Suppose the *government* passed a law limiting goods to *one use* or someone developed a good that instantly destroyed itself after the first use with no change in costs. It is easy to see that with our assumptions which include independent demand schedules such a law would decrease, and could result in a complete cessation of, sales of new units. New sales *would cease if C exceeded the intersection of* D_1 *and the price* axis in the figures. And profits would be less. Neither the monopolist in figure 2 with $X_1^* = X_2^*$ nor the monopolist in figure 3 with $X_1^* > X_2^*$ would want to kill off the secondhand market. If we assume that cost of production does vary with the life of the good the fall in costs attenuates the decrease in revenues. Transaction costs operate, apart from other considerations, to decrease the scope of secondhand markets.

Now D_1 and D_2 may be *independent* if they represent demand for diamonds by present and future generations or demands based on quite different uses of a good.[4] But in analyzing new and secondhand demands

---

[4] Demand for many items is approximately independent of demand for new items due to (relatively) noncompeting uses, e.g., demand for furnishings for summer cottages.

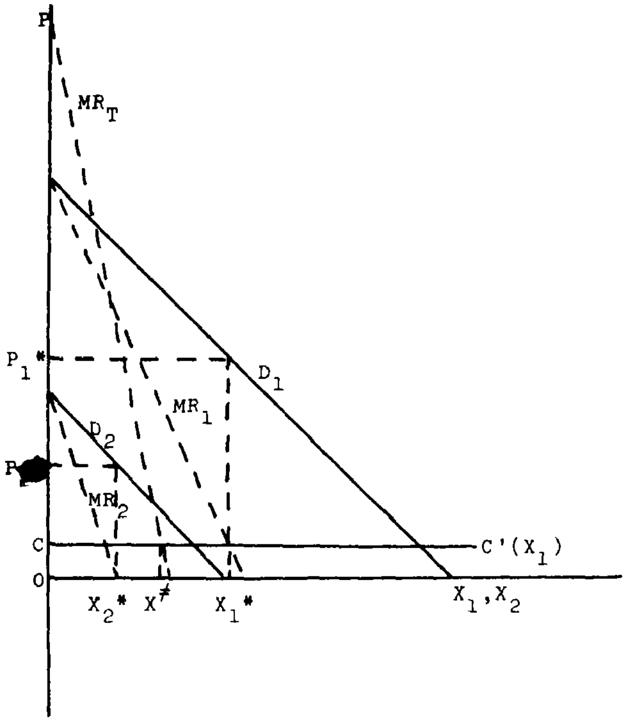


FIG. 3

in general, we cannot assume that D_1 and D_2 will be independent.[5] New and secondhand goods will seldom be perfect substitutes at identical prices. But they will usually be good substitutes, where $P_1 = P_1(X_1, X_2)$ and $P_2 = P_2(X_1, X_2)$ with $(\partial X_j/\partial P_i) < 0$, $i = j$, and $0 < (\partial X_j/\partial P_i) < \infty$, $i \neq j$.

II

It is not difficult to determine the effect of interdependent demands so long as one assumes that it pays to sell the same number of units in each market, that is, that $X_1 = X_2$. It can be shown that a monopolist maximizes his profits in this case by equating marginal cost, $C'(X_1)$, and total marginal revenue, $MR_T = (\varepsilon_1 + \alpha)P_1 + (\varepsilon_2 + \beta)P_2$ where ε_1,

---

[5] Much of the interesting work on the peak pricing problem assumes independent demands (see, e.g., Littlechild 1970). A brief summary of that paper appears in the discussion of the Kuhn-Tucker theorem in Baumol (1972, pp. 167–69).

ε_2, α, and β are functions of the various elasticities of demand

$$\varepsilon_1 = 1 + \frac{X_1\,\delta P_1}{P_1\,\delta X_1}, \qquad \varepsilon_2 = 1 + \frac{X_2\,\delta P_2}{P_2\,\delta X_2},$$

$$\alpha = \frac{X_2\,\delta P_1}{P_1\,\delta X_2}, \qquad \beta = \frac{X_1\,\delta P_2}{P_2\,\delta X_1}$$

An approximate graphical representation would be obtained by appropriate substitution of these expressions for the marginal revenue functions in figure 2 These elasticities are functions of X_i and P_i such that (X_j/P_i) $(\partial P_i/\partial X_j)$ need not equal nor have the same sign as $1/[(P_i/X_j)$ $(\partial X_j/\partial P_i)]$ [6] One can establish the conditions for them to have the same sign but can say nothing about any actual situation without quantitative values The interdependent case reduces to the independent case if $\alpha = \beta = 0$

Again, it might occur that $MR_2 < 0$ with $X_1 = X_2$, and the rule is that marginal cost must equal total marginal revenue, $MR_1 \geq 0$, $MR_2 \geq 0$ And a monopolist who sells his product will buy back some units to achieve the result otherwise attained by renting

What about a law limiting a product to one use or development of a product that destroys itself after one use with no change in costs? These would be like an infinite increase in the price of used books and would shift demand for new books upward Whether profits would increase or decrease--the limiting case would again be that production would cease entirely - would depend upon the substitutability of new and used books and the effect on costs of decreased durability Although this argument is formally correct, it does not lend itself to clear-cut answers A slightly different way of viewing the problem may shed added light on the possible outcomes

Suppose that new and used books are perfect substitutes, that no resale is permitted, and that demand for new books is given by the linear demand function, D, in figure 4 If resale is now permitted, one-half of the customers will by assumption shift to used books, and demand for new books will change so that the price will be doubled for halved quantities [7] New-book buyers will pay twice what they paid before, D', in anticipation of receiving one-half that price as a trade-in value 1 year later The net price of new and used books would be equal If costs of production were zero, output of new books would be one-half of what

---

[6] $(\partial P_i/\partial X_j)$ and $(\partial X_j/\partial P_i)$ have the same sign in $P_i = P_i(X_i, X_j)$, but $(\partial P_i/\partial X_j)$ in $P_i = P_i(X_i, X_j)$ and $(\partial X_j/\partial P_i)$ in $X_j = X_j(P_i, P_j)$ are derivatives of different functions and may not have the same sign

[7] Jack Hirshleifer suggested in correspondence that this might be a fruitful way to approach the problem

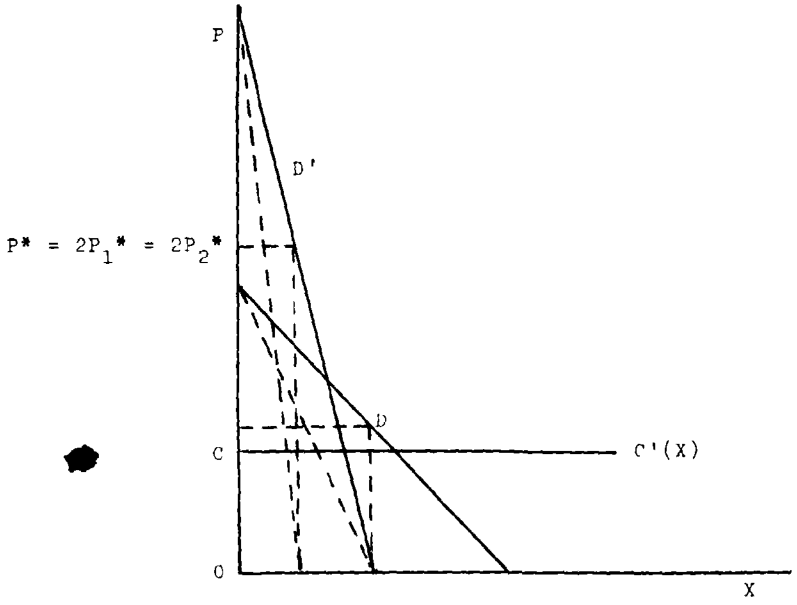


FIG. 4

it was before, the price of a new book would be $P^* = 2P_1^* = 2P_2^*$, and total revenue and profits would be unchanged by introduction of a secondhand market.

A product that destroys itself after one use would involve the reverse movement from *D′* to *D*. If costs were zero, output would double, and total revenues and profits would be unchanged by killing off the used market.

With positive costs of production, output would less than double, and profits would be lower without a secondhand market. When costs fall significantly with durability, these effects are attenuated. Although costs "could" fall by more than one-half when the life of a good is halved, a rational publisher not subject to transaction costs would never choose to produce the more durable book in the first place.[8]

Used books are not in fact perfect substitutes for new ones, so *D′* in the more general case will be below *D′* in figure 4. Introduction of a secondhand market would not double the demand price at halved quantities Killing off the market would decrease demand price by less than one-half at doubled quantities It is conceivable—though hardly necessary—that some combination of revenues and costs would make it pay to eliminate the secondhand market in accord with the publishers' view of the situation We have no information on relevant revenues and costs, however, that makes this probable The probability decreases with increases in the marginal rate of substitution between new and used books

---

[8] A simple example will support the statements in these three paragraphs. Suppose that new and used books are perfect substitutes and that demand with no resale permitted, *D*, is given by $P = 3 - X$, so demand with resale permitted, *D′*, is given by $P' = 6 - 4X'$. If $AC = MC = 0$, revenue and profit will be 2.25 in both cases. If now $AC = MC = 1$, profit with no resale and profit with resale will be 1.00 and 1.94, respectively. If costs increase as durability increases, $AC = MC = 2$ for a book with twice the life expectancy marks the point where profit could be increased by killing off the secondhand market.

There is no reason to suppose that we would obtain different results if we had assumed three or more uses Storage and disposal costs and transaction costs would add realism to the picture but are not essential in an examination of the chief issue

With regard to the quotation appearing in the first paragraph, a monopolist may or may not increase profits by changing his product A monopolistic competitor must, of course, keep the ideas and facts in his book up-to-date or find his demand curve shifting to the left It is not as obvious as the quotation suggests that the reason for bringing out a new edition is to kill off the existing used-book market

References

Baumol, William J *Economic Theory and Operations Analysis* 3d ed Englewood Cliffs, N J Prentice-Hall, 1972

Ferguson, Charles E *Microeconomic Theory* Rev ed Homewood, Ill Irwin, 1969

Friedman, Milton *Price Theory* Chicago Aldine, 1962

Hirshleifer, Jack "Suppression of Inventions" *J P E* 79 (March/April 1971) 382-83

Littlechild, Stephen C "Pricing Telephone Services" *Bell J Econ Management Sci* 1 (Autumn 1970) 191 210

Miller, H Laurence, Jr "A Note on Fox' Theory of Second-Hand Markets" *Economica*, n s 27 (August 1960) 249-52

A Note on Income (Labor) Inequality: Income Tax Systems and Human Capital Theory

Claude Montmarquette

Université de Montréal

Human capital theory has recently provided us with a theory of income distribution and an explanation of the income inequality existing in a given country at a specific point in time However, works in this area have generally neglected to consider the income tax structure with which a potential investor will be confronted when returns from his investment are to be received [1] On the other hand, measures of income inequality (specifically defined from a Lorenz curve) have been discussed and related to lump-sum and proportional taxes but outside the context of human capital theory (Levine and Singer 1970)

The purpose of this note is to investigate, with the help of human capital theory, the before- and after-tax income distributions associated with different income tax systems The discussion formalizes the widespread belief in the literature of income distribution (Aitchison and Brown 1957, Goode 1964) that the income tax structure influences the pattern of before-tax income I then analyze the problem of income inequality for the different income tax systems using the variance of the logarithm of income as a measure of this inequality

Let us follow Mincer (1958, 1970), Chiswick (1971), and others and assume that all schooling and training costs are viewed by the individual as time costs In equilibrium, individuals with larger amounts of training must be compensated for their costs In other words, neglecting ability and quality issues, we must equalize the expected present value of lifetime earnings at the time the training is undertaken Then, by assuming expectations to be realized,[2] we can explain observed income by past investments in human capital [3]

My thanks to Edi Karni for helpful comments on an earlier version of this notc Several of my colleagues, a referee, and Robert J Gordon also made useful comment Any errors are mine

[1] To mention a few exceptions, Hansen (1963) and Becker (1964, 1967)

[2] Note the distinction between anticipations on the return from schooling and th realized return of this investment The distinction respecifies the identification proble which relates schooling to income, and vice versa The distinction is also conceptual helpful in stressing the difference between the allocative effects of a tax and its distributi effects

[3] For simplicity of discussion, I am considering the schooling model, a simplifi version of the human capital model

Mathematically we have

$$V_s^A = Y_s^A \int_s^{n+s} e^{-rt}\,dt = Y_s^A e^{-rs}(1 - e^{-rn})/r, \qquad (1)$$

where V_s^A = expected present value of lifetime earnings of an individual with s years of schooling, Y_s^A = expected annual earnings of an individual with s years of schooling, r = the interest rate of the expected return on schooling, n = adult life to retirement, t = time, and e = base of natural logarithms

Similarly, for an individual with $s - d$ years of schooling we obtain (assuming the number of years spent working equal to n for all schooling groups)

$$V_{s-d}^A = Y_{s-d}^A \int_{s-d}^{n+s-d} e^{-rt}\,dt = Y_{s-d}^A e^{-r(s-d)}(1 - e^{-rn})/r \qquad (2)$$

In equilibrium, $V_s^A = V_{s-d}^A$, and we obtain $K_{s,s-d}^A$, the ratio of expected annual earnings of persons differing by d years of schooling Now assume the expectations to be realized, then $V_s = V_{s-d}$ will yield

$$K_{s,s-d} = \frac{Y_s}{Y_{s-d}} = \frac{e^{-r(s-d)}}{e^{-rs}}, \qquad (3)$$

the ratio of annual earnings of persons differing by d years of schooling Let us specify Y_s and Y_{s-d} as after-tax income By doing so we assume that individuals are concerned with the expected values of these variables in their choice of a specific income stream [4] Again, by focusing on the inverse relationship, realized after-tax income will be explained by investments in human capital [5] Before-tax incomes (Y^g) are related to after-tax incomes by the usual expressions

$$Y_s = Y_s^g(1 - T_s), \qquad (4)$$

$$Y_{s-d} = Y_{s-d}^g(1 - T_{s-d}), \qquad (5)$$

with T_s and T_{s-d} as the relevant average income tax rates [6]

---

[4] In a complete model, the individual will adjust to the distribution pattern (if one exists) of tax proceeds Note that the distribution of the proceeds need not be related to a specific income tax system

[5] Mincer (1970) did not distinguish between before- and after-tax income distribution The distinction has some relevance in testing the ability of the human capital theory to explain income distribution

[6] Tax rates are indexed with the schooling variable, as they depend indirectly on the level of schooling Note that formally

$$T_s = \sum_{s=0}^{m} \frac{t_s(Y_s - Y_{s-1})}{Y_s},$$

where t_s is the marginal tax rate of a given income bracket

Forming the ratio $K^*_{s,s-d}$ of before-tax annual incomes, we obtain[7]

$$K^*_{s\,s-d} = \frac{Y_s^g}{Y_{s-d}^g} = \frac{Y_s(1 - T_{s-d})}{Y_{s-d}(1 - T_s)} = K_{s,s-d}\,\frac{(1 - T_{s-d})}{(1 - T_s)}$$

$$= \frac{e^{-r(s-d)}(1 - T_{s-d})}{e^{-rs}(1 - T_s)} \tag{6}$$

The last two equalities are derived from (3). With $s - d = 0$ for simplicity,[8] equation (6) reduces to

$$K^*_{s\,0} = \frac{Y_s^g}{Y_0^g} = \frac{Y_s(1 - T_0)}{Y_0(1 - T_s)} = K_{s,0}\,\frac{(1 - T_0)}{(1 - T_s)} = e^{rs}\,\frac{(1 - T_0)}{(1 - T_s)} \tag{7}$$

With $T_s = T_0$ (proportional income tax), equation (7) gives Mincer's schooling model (1970). As he points out: "The formulation highlights the principle that it is not the shorter pay-off period of the more educated (recall that same span n of life time earnings is assumed) but the postponement of earnings that is the basic cause of differential in earnings" (Mincer 1970, p. 7). With $T_s \neq T_0$, another cause of differential in gross earnings is added.

From (7), we can then easily derive for different income tax systems the before- and after-tax income distributions as explained by the schooling model. Table 1 summarizes the results; it also gives the variance of the logarithm of income, a commonly used measure (hereafter called F-measure) of income inequality.[9]

From the table, it is easily seen that, in the case of a proportional income tax system, the same measure F of income inequality holds for both before- and after-tax income distribution.[10] Furthermore, if we assume for the three income tax systems the same rate of return on schooling and the same variance of schooling, the F-measure of the after-tax income inequality will be the same under any of the three income tax systems. Finally, under the same assumptions, the progressive income tax system will present, relative to the case of a proportional income tax system, a greater inequality in the before-tax income distribution. This is also true for the regressive income tax system, provided

---

[7] The same result could be obtained by forming the after-tax present value function

$$V^* = Y_s^g(1 - T_s)\int_s^n e^{-rt}\,dt$$

and by repeating the analysis. The presentation in the text explores the analogy with the treatment in investment literature of the physical capital affected by a corporate income tax (see Harberger 1962, 1968). Note also Friedman's (1962) similar discussion in terms of wage rate equalization rather than capital value equalization.

[8] The constant $Y_{s-d}^g = Y_0^g$ represents the income derived from initial endowment.

[9] For an appraisal of this measure of income inequality, see Weisskoff (1970).

[10] This result was also derived by Levine and Singer (1970) for measures of income inequality defined from the Lorenz curve.

TABLE 1

INCOME TAX SYSTEMS AND INCOME DISTRIBUTIONS

Before-Tax Income	After-Tax Income
Proportional Income Tax System ($T_s = T_0$)	
$Y^g_s = Y^g_0 e^{rs}$ $\text{var} (\log Y^g_s) = r^2 \text{ var} (s)$*	$Y_s = Y_0 e^{rs}$ $\text{var} (\log Y_s) = r^2 \text{ var} (s)$
Progressive Income Tax System ($T_0 = 0,\ T_s > 0$)†	
$Y^g_s = Y^g_0 e^{rs}/(1 - T_s)$ $\text{var} (\log Y^g_s) = r^2 \text{ var} (s)$ $+ \text{ var} [\log (1 - T_s)]$ $- 2r \text{ cov} [s, \log (1 - T_s)]$	$Y_s = Y_0 e^{rs}$ $\text{var} (\log Y_s) = r^2 \text{ var} (s)$
Regressive Income Tax System ($T_0 \neq 0,\ T_0 > T_s,\ \forall T_s$)	
$Y^g_s = Y^g_0 e^{rs}(1 - T_0)/(1 - T_s)$ $\text{var} (\log Y^g_s) = r^2 \text{ var} (s)$ $+ \text{ var} [\log (1 - T_s)]$ $- 2r \text{ cov} [s, \log (1 - T_s)]$	$Y_s = Y_0 e^{rs}$ $\text{var} (\log Y_s) = r^2 \text{ var} (s)$

* For simplicity r is taken as given

† Note that $K_{s,0} > 1$ so $T_0 = 0$ and $T_s > 0$ imply a progressive income tax structure, since the marginal tax rate increases with income (see n 6)

however that $2r \text{ cov} [s, \log (1 - T_s)] < \text{var} [\log (1 - T_s)]$ is satisfied

The presentation clearly suggests that a progressive income tax system might create inequalities in the before-tax income distribution that would counteract the mathematical effectiveness of this progressive taxation to reduce after-tax income inequality[11] More specifically, to the extent that investment in human capital alone explains income differences, and to the extent that returns to investment in education are equalized at the margin, income taxes are completely shifted forward in the long run This result is not original, since incidence theory maintains that income taxes are completely shifted forward, as long-run supply in any occupation can be considered perfectly elastic[12] The merit of this note lies in the derivation, in terms of a specific measure of income inequality, of sufficient conditions for this last result to hold Furthermore, these conditions appear to be sufficient conditions for the Pareto law to hold, as this law states that income is distributed in the same way regardless of social and political institutions and regardless of taxation The law implies that allocative and distributive effects of an income tax system need not be conflicting [13]

---

[11] The concept of progressivity in taxation is not trivial, since it must be defined for the tax system in general and not only for the income tax However, what I discuss in the text are the effects on income distribution of an effective progressive income tax

[12] A referee of this *Journal* drew this point to my attention

To discuss what effectively happens to those conditions, one must first investigate the allocative effects of the different forms of income tax on the investments in human capital Then it remains to evaluate the effects of a change (if any) in investments in human capital on the rate of return and on the variance of these investments The schooling model and the restrictive assumptions retained in this note might help answer these questions However, in the context of a general equilibrium framework of incidence theory (Harberger 1962, Mieszkowski 1967) and with a model of income inequality which introduces ability and quality considerations and refers to a general version of the human capital model, those questions appear quite complex and no attempt will be made here to discuss them

In this note, I have shown that the human capital model of income distribution enables us to incorporate the income tax system of a society into an analytical framework which explores the relationship between the before- and after-tax income distributions The importance of the results presented in the text is strictly bound to the importance of the human capital model to explain the difference in earnings, which, in the final analysis, proves to be an empirical question

References

Aitchison, J , and Brown, J A C *The Lognormal Distribution* Cambridge Cambridge Univ Press, 1957

Becker, G S *Human Capital* New York Columbia Univ Press, 1964

——— *Human Capital and the Personal Distribution of Income An Analytical Approach* Ann Arbor Univ Michigan Press, 1967

Chiswick, B R "Earnings Inequality and Economic Development " *Q J E* 85, no 1 (February 1971) 21–40

Friedman, Milton *Price Theory A Provisional Text* Chicago Aldine, 1962

Goode, R *The Individual Income Tax* Washington Brookings Inst , 1964

Hansen, W L "Total and Private Rates of Return to Investment in Schooling " *J P E* 71, no 2 (April 1963) 128–40

Harberger, A C "The Incidence of the Corporation Income Tax " *J P E* 70, no 3 (June 1962) 215–40

——— "Corporation Income Taxes " In *International Encyclopedia of the Social Sciences* Vol 2 New York Macmillan, 1968

---

[13] Note that as human capital models attempt to explain labor income inequality, equity issues are not a difficulty Furthermore, to extend the results of the text to the general human capital model to explain income distribution and income inequality (see Chiswick 1971), conditions similar to those affecting the return and variance of schooling must prevail on the return, variance, and covariance of all other forms of investment in human capital

Levine, D. B., and Singer, N M. "The Mathematical Relation between the Income Density Function and the Measurement of Income Inequality" *Econometrica* 38, no 2 (March 1970) 312–24

Mieszkowski, P "On the Theory of Tax Incidence" *J P E* 75, no 3 (June 1967): 215–40

Mincer, J "Investment in Human Capital" *J P E* 66, no 4 (August 1958) 281–302

——— "The Distribution of Labor Incomes A Survey with Special Reference to the Human Capital Approach" *J Econ Literature* 8, no 1 (March 1970) 1–26

Weisskoff, R "Income Distribution and Economic Growth in Puerto Rico, Argentina and Mexico" *Rev Income and Wealth* 16, no 4 (December 1970) 303–32

A Note on Equilibrium in Monopolistic Competition

Yoram Peles

The Hebrew University

Two major results obtained in Chamberlinian monopolistic competition are (*a*) in the long-run equilibrium the firm earns only normal profits, and (*b*) it has excess capacity, that is, the firm produces in the range of declining average costs. This paper emphasizes the importance of short-run fixed cost of product differentiating, such as store location, advertising, R & D, etc., for the determination of both the location of the demand curve and short-run equilibrium. Moreover, in the extreme case of completely immovable assets it will be shown that under certain conditions, the short-run equilibrium solution in which there exists both excess profits and rising average costs obtains in long-run equilibrium as well.

The excess-capacity theorem is based on the existence of a U-shaped cost curve. However, the declining portion of the long-run average cost curve may serve as a barrier for new firms. This barrier would arise if there is not enough room to support new firms at the prevailing price, though existing firms enjoy excess profits and not necessarily excess capacity.

An example sometimes given for the monopolistic competition is spatial competition—the location of firms along a line. For convenience let us analyze a specific example, however, the conclusions are generally valid for this type of competition. Let consumers be spread evenly along a road going from L to M. At every point along the road the demand for a given product is given by $P = B - q - aT$, where B is a constant, a denotes the consumer's cost of transporting one unit of good one unit of distance, T is the number of units of distance from point of selling to point of destination, and q is quantity demanded at a particular point on the road. Put $a = 1$. Information is perfect and available to all, hence consumers do not shop around but buy from the seller whose price plus cost of transportation is the lowest. All producers have the same cost function, which includes an initial investment in fixed assets

---

I acknowledge the helpful comments of Lester Telser and a referee.

(or better with no alternative use in the short run) denoted F, and zero marginal costs. The annual cost of capital is α Let us start by assuming that these fixed assets are not amortized Firms are free to enter the market and they will do so as long as it pays them, namely, if they expect annual revenue $\geq \alpha F$ Firms are free to change the price of their product and prices may differ between producers, but it is not possible to have price discrimination by the same producer

Now, firms start to enter this market If there is only one producer along the line he will charge a monopoly price P which gives the maximum revenue In this example, the total demand confronting the producer, to the left and right together, is given by

$$Q = 2\int_{T=0}^{B-P} (B - P - T)dT$$

The price charged is therefore $\frac{1}{3}B$, selling to customers up to a distance of $\frac{2}{3}B$ to his left and $\frac{2}{3}B$ to his right (fig 1) The producer might locate himself at point $\frac{2}{3}B$ thereby securing his left wing from any new entry, as will be shown At this point he sells a quantity $q = B - \frac{1}{3}B = \frac{2}{3}B$, to his left, he sells up to the beginning of the road at point L, at which point $q = 0$, to his right he sells up to point $1\frac{1}{3}B$, where again $q = 0$ Total quantity sold is $\frac{4}{9}B^2$ units and $PQ = \frac{4}{27}B^3$ A second producer will sell under the same conditions, so let him be located at point $2B$, once again securing his left wing from new entry, with a market range from point $1\frac{1}{3}B$ to $2\frac{2}{3}B$ So, as long as there is room along the road toward M, new producers will enter on these terms

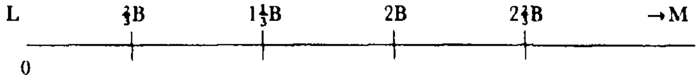


Fig 1

Suppose the road has already been colonized up to M and a new potential producer wants to enter the market, knowing that firms are already located at points $\frac{2}{3}B$, $2B$, etc , and they cannot move away (short-run analysis) The best place for locating his firm is at the midpoint between two producers, such as at point $1\frac{1}{3}B$ The intruder enters the market at this point Now if prices are still kept at $P = \frac{B}{3}$, the intruder will take away customers from points B to $1\frac{2}{3}B$, having $PQ = \frac{3}{27}B^3$ However, once the intruder is in the market, $P = \frac{B}{3}$ is not the best price for either the neighboring old producers or the new one With a new producer located at midpoint between two old ones, the best policy for the existing firms and the new ones together is to charge still the same price.[1] Hence each producer sells a distance of $\frac{B}{3}$ on either side, therefore having a demand of

$$Q = 2\int_{T=0}^{B/3} (B - P - T)dT$$

The new price set is $P = 1\ 25\frac{B}{3}$, at which the entrant sells a quantity of $Q = \frac{2\ 5}{9}B^2$, getting a revenue of $\frac{3\ 125}{27}B^3$

So, as long as the old producers do not change their location, the potential producer cannot expect an annual profit of more than $\frac{3\ 125}{27}B^3$ With cost of capital α he will not enter the market as long as $F > \frac{3\ 125}{27}B^3\alpha$, although those already in the market might benefit from monopoly profits (if $F < \frac{B^3}{27}\alpha$)

If the assets are amortized over a finite number of years, the time will come when they have to be replaced (long-run analysis) Then it will be to the benefit of the old producers to move away from a new intruder, if he enters the market, thus increasing his market size and allowing him more profi[illegible] than the maximum $\frac{3\ 125}{27}B^3$ given above In general the longer is this long-run the lower is the probability of a new entry and therefore the stronger the monopoly conditions for those already in market

Let $F < \frac{3\ 125}{27}B^3\alpha$, the old producers can adopt a policy of preventing new entry and still getting a monopoly profit In the previous case each producer kept the best market interval giving him the maximum monopoly profit If producers expect a new entry, they might locate themselves closer together from the beginning, reducing the market area and the maximum profit an intruder can get The cost of this strategy for the old producers is a lower monopoly profit[2]

In the previous case each firm was producing under conditions of constant short-run average costs but declining long-run ones However, having a long-run analysis, in which long-run costs include both short-run variable costs and short-run fixed costs, the monopolistic competition solution might exist at the decreasing, minimum, or increasing part of the long-run average cost curve (and, as before, with excess profits) To show this let us take the extreme demand condition of zero price elasticity below a given price at every point along the road but positive elasticity of demand with respect to distance Let the demand at every point along the road be $q = 1$ unit of product at $P \leq B$ and $q = 0$ at $P > B$ The first producer entering this market can sell up to a distance of $T = B - P$ to the right and the same distance to the left The quantity demanded to his right and left together is the quantity demanded at each point q times the distance, that is $Q = q \times 2T = q \times 2(B - P)$, and since $q = 1$, therefore $Q = 2B - 2P$ or $P = B - \frac{Q}{2}$. In order to simplify the mathematical expressions (though not affecting the conclusions) let $B = 100$, such that $P = 100 - \frac{Q}{2}$. Let short-run $mc = Q^2$, then the best price for the first firm to enter the market is $P = 95.25$ with $Q = 9.5$ and $T = 4.75$. The total revenue $PQ = 904.9$, and total short-run variable costs are 285.8. This solution is at the minimum long-run average cost if this minimum cost amounts to 90.5, so that the annual outlay on the short-run fixed cost is 574. Now, after the road is colonized with firms located at intervals of 9.5 starting at point 4.75, a new firm wanting to enter this market can expect the best at the midpoint between two old producers, with a market range of 4.75. With such range the highest profit is obtained charging $P = 97.625$, selling $Q = 4.75$, with revenue $PQ = 463.7$, and after subtracting 35.7 for short-run variable cost, the potential firm can expect an income of no more than 428, which falls short of the annual cost for short-run fixed cost of 574.

---

[1] The following analysis assumes that between every two old producers enters a new one If only one new producer enters the market, the producers already in the market prefer a lower price than that of the intruder because they face competition only on one side of the road However, even with the higher price the conclusions will not be changed

[2] See Pashigian 1968

If no new entry occurred by the date of replacement, each firm—whether already in market or a potential intruder, will try to be the first to build the new plant. Since everyone knows what is going to happen on the day of replacement, the potential intruder (and therefore old firms already in business) may anticipate the day of replacement by building the new plant earlier. It will pay them to do so as long as the expected future monopoly profits exceed expected losses until the old plant is scrapped. In other words, firms will obtain normal profits only from the new plants. The new plant can be sited anywhere along the road from a point immediately next to an old plant to a point midway between two old plants. Once the old plant is scrapped,[3] the best location for the new plant is at the site of the old, with the new plant serving the road interval previously served by the old plant and keeping the old prices and quantities. The excess profit now pays for the premature investment.[4]

However, while the old plant is still in use, short-run profits are at a maximum when the new plant is located at the midpoint between two old plants. Hence, a new plant built before an old one is scrapped will be somewhere between the site of the old plant and the midpoint—closer to the midpoint the further ahead is the day of scrapping. It is better to stop at this point before the picture of the very long-run solution becomes too complicated.

---

[3] If the new plant is owned by an old producer, early investment can be accompanied by premature scrapping.

[4] Instead of building a new plant early the entrepreneur may declare his fixed intention of building the new plant at the due time and so save losses. However, let us not go too far into strategies of this kind.

In short, monopolistic competition can operate under conditions where each firm faces competition only from neighboring firms, whether in terms of distance or quality The existence of untransferable fixed assets allows individual firms to adopt a policy leaving them monopoly conditions with excess profits, although free and fast entry is allowed The quantity these firms produce might be at a minimum point on the average cost curve (or along the increasing part of the curve) and not necessarily at decreasing average cost, as usually claimed Hence firms can produce "efficiently "[5]

References

Pashigian, B Peter "Limit Price and the Market Share of the Leading Firm " *J Indus Econ* 16 (July 1968) 165–77

Telser, Lester G "On the Regulation of Industry A Note " *J P E* 77 (November/December 1969) 939–52

---

[5] On the problem of optimum social production under the above conditions of spatial competition, see Telser (1969)

Optimal Replacement of Capital Goods in Early New England and British Textile Firms: A Comment

David Denslow, Jr., and David Schulze

University of Florida

Professor Jeffrey G. Williamson (1971) states in a recent article that if all technical progress is labor saving and capital embodied then "the level of unskilled wages prevailing is irrelevant" (p. 1327) to the optimal frequency of replacement of capital equipment, which is contrary to what Habakkuk (1962) has written. "It appears," says Williamson, "that Habakkuk's model is not internally consistent" (p. 1327).

In fact, however, Williamson's model is inconsistent.[1] His is a vintage model in which a physical unit of capital, $K(v)$, is operated by a homogeneous labor input, $L(v)$, to produce textile output, $X(v)$, from machinery of vintage v. The capital-output ratio, $\alpha^{-1} = K(v)/X(v)$, is assumed constant over time, while the capital-labor ratio, $b(v) = K(v)/L(v)$, is assumed to increase exponentially

$$b(t) = e^{\mu(t-v)}b(v) \tag{1}$$

Because of the labor-saving technical progress expressed by (1), firms expect commodity prices to fall at the same rate as $b(t)$ increases

$$P(t) = e^{-\mu(t-v)}P(v) \tag{2}$$

On these assumptions it may be shown that the present value at time v of net revenue over the entire useful life, u, of a machine of vintage v is

$$\int_{v}^{v+u} [\alpha P(v)e^{-\mu(t-v)} - wb(v)^{-1}]e^{-i(t-v)}\,dt,$$

where w is the annual wage rate for unskilled labor and i is the money rate of interest. The firm assumes both i and w to be constant. The net worth of the new acquisition can be written as the difference between

---

The authors are indebted to Rafael Lusky for helpful discussions.

[1] We have borrowed freely from Williamson's phrasing in the following summary of his model.

this net income stream and the price, p, of a new capital good of vintage v:

$$n(v) = \frac{\alpha P(v)}{\mu + i}[1 - e^{-u(\mu+i)}] - \frac{w}{ib(v)}(1 - e^{-iu}) - p \qquad (3)$$

The price, p, of capital goods is assumed to remain constant over time

Williamson says that "by maintaining *constant* capacity and replacing the capital stock every u years, we can get an expression, $A(0)$, for the net worth of future physical capital stocks required for replacement" (p 1325) [2] He derives an equation for $A(0)$ as follows "The physical stock of capital required for the jth replacement is $K(ju) = x(v)/\alpha = X(0)/\alpha$, and its net worth at $t = 0$ is $\alpha^{-1}X(0)n(0)(1 + e^{-iu} + e^{-i2u} + + e^{-iju} +)$ Thus $A(0) = [X(0)n(0)]/[\alpha(1 - e^{-iu})]$" (p 1325, n 5)

Williamson rewrites the above equation as

$$A(0) = \frac{N(0)P(0)^{\eta}}{\alpha}\left\{\frac{\alpha P(0)}{\mu + i}\frac{[1 - e^{-u(\mu+i)}]}{(1 - e^{-iu})} - \frac{w}{ib(0)} - \frac{p}{(1 - e^{-iu})}\right\} \qquad (4)$$

To obtain the first-order condition for maximization of net worth, he sets $dA(0)/du = 0$ and rearranges

$$\frac{pi}{\alpha P(0)} = \frac{1}{\mu + i}[i + \mu e^{-u(\mu+i)} - (\mu + i)e^{-u\mu}] \qquad (5)$$

Equation (5) presents a couple of puzzles First, one would not expect the initial price of output to affect optimal replacement, since Williamson writes that "our interest is in the replacement decision under constant capacity conditions" (p 1325) But from implicit differentiation of (5),

$$\frac{du}{dP(0)} = \frac{ipe^{\mu u}}{\alpha\mu P(0)^2(e^{-iu} - 1)}, \qquad (6)$$

which is negative The frequency of replacement would increase with a rise in the initial price level, even though at constant capacity the frequency of replacement can have no effect on total sales revenue To put it another way, under conditions of arbitrarily selected constant

---

[2] Italics in the original Williamson says his model is an extension of one presented by Brems (1968) Brems's model is one of neutral technical progress None of our criticisms of Williamson's model applies to Brems's Williamson's appendix follows Brems's presentation closely and contains minor errors only when Williamson deviates from Brems The equilibrium condition, for example, should be

$$\left(\frac{\mu}{\mu - i}\right)e^{-iu} - \left(\frac{i}{\mu - i}\right)e^{-\mu u} = \frac{ip}{w} + 1,$$

the equations are invalid when $\mu = 0$, and it is unreasonable to expect

$$i^{-1}\ln\left[\left(\frac{ip}{w}\right)\left(\frac{\mu - i}{\mu}\right)\right]$$

to assume negative values if that expression is a good approximation to u

capacity, one would expect maximization of net worth to have a cost minimization dual in which the price of output has no direct effect

Second, since newer machines use less labor than older ones to produce the same amount of output, one would expect that the higher the wage level, the more frequently the entrepreneur would replace machines But in Williamson's model, as he points out, "unskilled wages do not appear to influence the replacement decision at all" (p 1327)

Such implications lead one to suspect that equation (4) is incorrect, and indeed it is To see why, let us spell out the implicit steps in Williamson's derivation of that equation His crucial sentence is, once more "The physical stock of capital required for the jth replacement is $K(ju) = X(v)/\alpha = X(0)/\alpha$, and its net worth at $t = 0$ is $\alpha^{-1}X(0)n(0)e^{-iju}$" (p 1325, n 5) Now the net worth at time ju of the capital required for the jth replacement is $\alpha^{-1}X(0)n(ju)$ To find the value of this net worth at $t = 0$, one must discount $\alpha^{-1}X(0)n(ju)$ by a factor of e^{-iju}, obtaining $\alpha^{-1}X(0)n(ju)e^{-iju}$ An implicit step in Williamson's derivation of equation (4), then, is

$$\alpha^{-1}X(0)n(ju)e^{-iju} = \alpha^{-1}X(0)n(0)e^{-iju} \tag{7}$$

or

$$n(ju) = n(0) \tag{7'}$$

Equation (7′) is incorrect, however, except for special cases irrelevant to Williamson's presentation To show this, we modify equation (1) to express $b(ju)$ in terms of $b(0)$ and equation (2) to express $P(ju)$ in terms of $P(0)$

$$b(ju) = e^{\mu ju}b(0), \tag{1'}$$

$$P(ju) = e^{-\mu ju}P(0) \tag{2'}$$

Substituting from equations (1′) and (2′) into equation (3), we obtain

$$n(ju) = \frac{\alpha P(0)e^{-\mu ju}}{\mu + i}[1 - e^{-u(\mu + i)}] \tag{3'}$$

Also, $n(0)$ may be expressed as

$$n(0) = \frac{\alpha P(0)}{\mu + i}[1 - e^{-u(\mu + i)}] - \frac{w}{ib(0)}(1 - e^{-iu}) - p \tag{3''}$$

Given equations (3′) and (3″), equation (7′) implies that

$$\left[\frac{\alpha P(0)}{\mu + i}(1 - e^{-u(\mu + i)}) - \frac{w}{ib(0)}(1 - e^{-iu})\right]e^{-\mu ju} = \frac{\alpha P(0)}{\mu + i}(1 - e^{-u(\mu + i)}) - \frac{w}{ib(0)}(1 - e^{-iu}) \tag{8}$$

or

$$[n(0) + p]e^{-\mu ju} = n(0) + p \tag{8'}$$

From equation (8′), either $n(0) = -p$ or $\mu = 0$ when ju is positive In the first case, the entrepreneur would not install the first machine The second case is not useful for analyzing the effects of technical progress Equation (4), which forms the basis of all of Williamson's article but the appendix, is thus seen to be wrong Given the assumption that an entrepreneur would be willing to replace his machinery every u years over an infinite horizon the correct form of equation (4) is

$$A(0) = \frac{X(0)}{\alpha} \sum_{j=0}^{\infty} n(ju)e^{-iju} \tag{9}$$

or

$$A(0) = \alpha^{-1}X(0) \left\{ \frac{\alpha P(0)}{\mu + i} - \frac{w}{ib(0)} \frac{(1 - e^{-iu})}{(1 - e^{-u(\mu + i)})} - \frac{p}{1 - e^{-iu}} \right\}$$

The first-order condition for maximization of net worth is

$$\frac{p}{u} = \frac{(1 - e^{-iu})^2}{i^2 b(0)(1 - e^{-(\mu + i)u})^2} \times [i(1 - e^{-(\mu + i)u}) - (\mu + i)(1 - e^{-iu})e^{-\mu u}] \tag{10}$$

At this point in our analysis assume we were interrupted by a textile manufacturer who happened to be in our office seeking advice about when to replace his machinery He would not obey equation (10), because his rule of thumb is never to invest in a machine costing p even if he plans to use it forever (to say nothing of only u years) unless the net annual return from it will be at least ip Equation (10) requires him to buy an infinite number of machines which could never repay their cost He quickly drew a sketch for us showing why (fig 1)

With dollars plotted against time, the upper line shows the flow of revenue, $\alpha P(t)$, from a single machine at time t, ignoring the elasticity of demand for simplicity,[3] and the lower line shows the flow of wages paid the labor that operates a machine of vintage t The vertical distance between the lines represents the flow of net revenue, $r(t, t)$, from a machine of vintage t at the instant it is installed The equation for that distance is

$$r(t, t) = \alpha P(t) - wb(t)^{-1} \tag{11}$$

or

$$r(t, t) = [\alpha P(0) - wb(0)^{-1}]e^{-\mu t}$$

Clearly, $r(t, t)$ approaches zero as t approaches infinity, and there is some t^* after which $r(t, t)$ is always less than ip

Chastened by the manufacturer's accusation that he would be the

---

[3] Brems included a parameter for the elasticity of demand because he was analyzing price determination as well as optimal replacement We do not see why Williamson included demand parameters

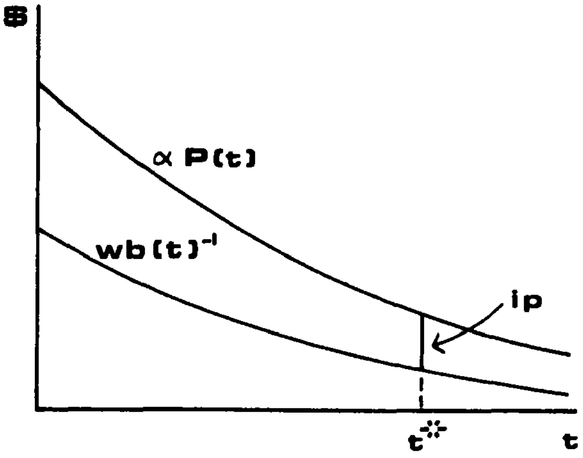


Fig 1

victim of an unreasonable requirement, we allowed him to change his price expectations so that $P(t)$ remains high enough to permit him to pay his wage bill and buy new machines each u years With that out of the way, we pointed out that product prices are irrelevant when the entrepreneur is constrained to constant output Equation (10) can be derived through minimization of the present value, $C(0)$, of costs per α^{-1} units of output per year

$$C(0) = \frac{w}{ib(0)} \frac{(1 - e^{-iu})}{(1 - e^{-(\mu + i)u})} + \frac{p}{1 - e^{-iu}} \tag{12}$$

As we were about to write out demonstrations that assuming second-order conditions hold, implicit differentiation of (10) yields du/dw negative, du/dp positive, and so on, the manufacturer broke in again "The gain from replacing a machine," he told us "is the saving in the wage bill from using less labor But since the wage bill approaches zero as a result of labor-saving technical progress and constant wage rates, there will come a time when I shall refuse to buy a new machine in order to reduce the wage bill" Since this argument was analogous to his earlier one, he did not have to draw us a picture 'I shall replace a machine," he continued, "whenever ip, the marginal cost of replacing it this instant instead of next, equals the reduction in the wage bill from installing a new machine."

Seeing how crestfallen we were at his forsaking an infinite horizon, however, he suggested a way out "My experience has been," he said, "that wages rise just as fast as new machines save labor" We realized that this was the right assumption to introduce stationarity into the problem, and that Williamson's mistake in adapting from Brems (1968) was to introduce nonstationarity when he assumed labor saving instead of neutral technical change[4]

Now let the entrepreneur's expectations about wages be

$$w(t) = w(0)e^{\mu t} \tag{13}$$

Equation (12) must be rewritten to accord with this new assumption

$$C(0) = \frac{w(0)(e^{\mu u} - 1)}{b(0)\mu(1 - e^{-iu})} + \frac{p}{1 - e^{-iu}} \tag{14}$$

The first-order condition for cost minimization is

$$b\mu i p w(0)^{-1} = \mu e^{(\mu + i)u} - (\mu + i)e^{\mu u} + i \tag{15}$$

Letting C_{uu} represent the second derivative of $C(0)$ with respect to u, we assume the second-order condition for cost minimization holds, $C_{uu} > 0$ From implicit differentiation of the first-order condition,

$$\frac{du}{dw(0)} = -(C_{uu}b(0)\mu)^{-1}(1 - e^{-iu})^{-2}[(1 - e^{-iu})\mu e^{\mu u} - (e^{\mu u} - 1)ie^{-iu}], \tag{16}$$

which is negative The higher the initial wage level, the greater the frequency of replacement

Similarly,

$$\frac{du}{dp} = \frac{ie^{-iu}}{C_{uu}(1 - e^{-iu})^2}, \tag{17}$$

which is positive The higher the price of machines, the lower the frequency of replacement The effect of μ on u is given by

$$\frac{du}{d\mu} = \frac{-w}{C_{uu}b(0)(1 - e^{-iu})}\left[\mu - \frac{ie^{-iu}}{(1 - e^{-iu})}\left(e^{\mu u} - \frac{(e^{\mu u} - 1)}{\mu^2}\right)\right] \tag{18}$$

The higher the rate of technical progress, the lower the frequency of replacement, since $du/d\mu$ is negative for plausible values of μ and i, specifically for

$$e^{\mu u}\left(1 + \frac{1}{i}\right) > 2 \tag{19}$$

It must be remembered, however, that an increase in μ implies not only more rapid technical progress but also more rapidly rising wages

---

[4] A model exhibits stationarity when historical time is irrelevant to inferences drawn from it Only relative time matters Our assumption that the rate of increase of wages equals the rate of technical progress is not satisfying but may be no more restrictive than Williamson's assumption (with which he was not content) that wages are constant Also, our revision of the model requires less restrictive assumptions about expectations regarding product prices than does Williamson's version

Equation (20) gives the effect of changes in interest rates on equipment longevity.

$$\frac{du}{di} = e^{-iu}(1 - e^{-iu})^{-3}C_{uu}^{-1} \times [i^2e^{-iu} + (1 - e^{-iu})(1 - i^2) + 2i^2e^{-iu}] \qquad (20)$$

The value of du/di is positive provided

$$i^2 + e^{-iu} < 1 \qquad (21)$$

From the preceding analysis it can be seen that the major conclusions of Williamson's article are unfounded He has not shown that with labor-saving capital-embodied technical progress "the level of unskilled wages prevailing in the textile sector is irrelevant" (p 1327) He has not shown that Habakkuk's model is internally inconsistent He has not shown that "a tariff structure which distorts relative prices by raising the relative price of consumption goods should increase the optimal replacement rate" (p 1330)

Why were there differences between American and British behavior in the replacement of textile machinery? More data about wages and machinery prices are needed before Habakkuk's hypothesis about the effect of relative wages on replacement can be tested Our analysis suggests that relative wages might matter Aside from that, Williamson's explanations at the beginning of his second section (pp 1321–22) appear plausible

References

Brems, Hans *Quantitative Economic Theory* New York Wiley, 1968

Habakkuk, H J *American and British Technology in the Nineteenth Century* New York Cambridge Univ Press, 1962

Williamson, J G "Optimal Replacement of Capital Goods The Early New England and British Textile Firm" *J P E* 79, no 6 (November/December 1971) 1320–34

Note Added in Proof The sentence following equation (18) should read "The higher the rate of technical progress, the higher the frequency of replacement, "

D D and D S

Optimal Replacement of Capital Goods in Early New England and British Textile Firms: Reply

Jeffrey G Williamson

University of Wisconsin

While attempting to make sense out of an important difference in early nineteenth-century Anglo-American replacement behavior, my 1971 article (Williamson 1971) developed a model which produced an apparent paradox The unskilled wage influenced replacement only under conditions of neutrality in technical progress In the historically more relevant case of labor-saving technical progress, it appeared that the unskilled wage played no role in determining the replacement decision in textiles Professors Denslow and Schulze correctly point out that the paradox was more apparent than real The first-order condition for the maximization of net worth derived from my original paper (p 1326) was

$$\left(\frac{K}{X}\right)^{*} = \frac{1}{iB}\left[i + \mu e^{-u(\mu+i)} - (\mu + i)e^{-u\mu}\right], \qquad (1)$$

where $B = (\mu + i)$ In the Denslow-Schulze version we get

$$\frac{p}{w} = \frac{(1 - e^{-iu})^{2}}{i^{2}b(0)(1 - e^{-(\mu+i)u})^{2}} \times \left[i(1 - e^{-(\mu+i)u}) - (\mu + i)(1 - e^{-iu})e^{-\mu u}\right], \qquad (2)$$

which looks very different indeed The differences rapidly diminish when the Denslow-Schulte expression (2) is rewritten

$$\left(\frac{p}{w}\right)b(0) = \frac{1}{iA}\left[i + \mu e^{-u(\mu+i)} - (\mu + i)e^{-u\mu}\right],$$

where $A = (1 - e^{-iu})^{-2}[i(1 - e^{-(\mu+i)u})^{2}]$ Recalling that $b(0)$ is the initial capital-labor ratio and multiplying numerator and denominator of the left-hand side by $P(0)X(0)$, we get a new corrected expression (3)

$$\left(\frac{K}{X}\right)^{*}\alpha_{L}(0)^{-1} = \frac{1}{iA}\left[i + \mu e^{-u(\mu+i)} - (\mu + i)e^{-u\mu}\right], \qquad (3)$$

where $\alpha_L(0)$ is the initial labor share in textiles. Expression (3) begins to look more like my original expression (1). Furthermore, it is far more useful for historical analysis, since income shares and capital-output variables are directly observable for the antebellum period, while input price ratios are far more difficult to document.

After having made this correction, Denslow and Schulze tend toward the overzealous. They may be forgiven, I suppose, since they apparently have no interest in the historical issues for which the analysis was used as a tool. They spend the remainder of their comment exploring the implications of their corrected model *under the assumption that entrepreneurs expect wages to rise at the same rate at which labor-saving technical change takes place.* One can, of course, assume anything one wishes, but in historical analysis one is compelled to be more discreet. My primary focus in the earlier paper was the 1820s and 1830s, when Anglo-American differences in replacement policy were most apparent. Scraps of real wage data from this period suggest very moderate, if any, rises in w. Lebergott (1960, p. 493) estimates that real wages of nonfarm employees fell by as much as 10 percent from 1818 to 1830 and rose by only 10–20 percent in the 1830s. Average real daily earnings of cotton textile workers actually fell by 10 percent between 1832 and 1840.[1] The evidence is no different regarding the ratio P/w. The price of machine-shop products declines from an index of 100 to 97 between 1834 and 1844. Money wages paid female operatives at Lowell after 1835 declined to 94 by 1842 but rose to 102 by 1846.[2] All of this evidence indicates that the Denslow-Schulze assumption makes their subsequent analysis a sterile exercise. Indeed, their inevitable result that higher rates of obsolescence (higher μ) implies lower rates of replacement is downright silly.

But let me return to the key issues in the debate. Denslow and Schulze conclude that when the correct expression (3) is employed in the historical analysis "the major conclusions of Williamson's article are unfounded." This is a curious statement, since it seems to me remarkable how little the conclusions of the earlier argument require revision!

Numerical analysis of equation (3) for $(K/Y)^* = 3.0$ and $\alpha_L = 0.403$, a capital-output ratio and labor share prevailing in American textiles in the 1830s, yields results very similar to those reported in figure 1 (Williamson 1971, p. 1329) of my original article. Thus, my earlier conclusion remains unchanged. "The evidence of far higher replacement rates in the United States textile sector in the 1820s and fairly similar replacement rates in the 1850s can only be explained by a convergence of expected rates of technical change between the two economies. . Interest rate differentials apparently had the opposite impact" (p 1328) Note further that $(K/X)^*\alpha_L(0)^{-1} = (p/w)b(0)$. Given b, the capital-labor ratio, constant, numerical analysis shows that a decline in p/w would induce lower replacement rates or a higher average age of equipment Since this key input price ratio does in fact decline during the 1830s and 1840s, we have an explanation for the convergence in Anglo-American replacement rates over the antebellum period In addition, my original conclusion that "the low relative price of investment goods created by the United States tariff structure in the 1820s should have further widened differences in replacement policy" (p 1331) still stands if we are willing to accept the conventional view that the tariff structure acted to raise the relative price of (industrial) wage goods, as it surely did [3]

---

[1] Money wages for U.S. cotton textiles from Lebergott (1960, p. 476, table 4) deflated by the Warren-Pearson price index.

[2] The prices of capital goods are from Brady (1966, p. 110, table 2a). The Lowell daily wage rates are taken from Lebergott (1960, p. 476, table 4).

References

Brady, D S.[illegible]rice Deflators for Final Product Estimates" In *Output, Employment, and Productivity in the United States after 1800* New York Nat Bur Econ Res, 1966

Brito, D L, and Williamson, J G "Skilled Labor and Nineteenth Century Anglo-American Managerial Behavior" *Explorations Econ History* 10, no 3 (Spring 1973) 235-53

Lebergott, S "Wage Trends, 1800-1900" In *Trends in the American Economy in the Nineteenth Century* Princeton, N J Nat Bur Econ Res, 1960

Williamson, Jeffrey G "Optimal Replacement of Capital Goods The Early New England and British Textile Firm" *J P E* 79, no 6 (November/December 1971) 1320-34

---

[3] A far more comprehensive analysis of contrasting Anglo-American nineteenth-century industrial managerial behavior can be found in Brito and Williamson (1973)

A Note on the Stability of Metzler's Phase Diagram

Don Patinkin

The Hebrew University

Lloyd Metzler's classic article on "Wealth, Saving, and the Rate of Interest"—which appeared in this *Journal* over 20 years ago—is noted for its clear and searching analysis of the implications of the Pigou or real-balance effect, and for the crucial distinction it made in this context (from the viewpoint of optimum-portfolio-composition, and hence neutrality) between a monetary change generated by a current deficit and one generated by an open-market operation

But in addition to this valuable substantive contribution, Metzler's article also made an important contribution to technique by providing a graphical analysis of the stability of the system in terms of what is now termed a "phase diagram " And though such a diagram is essentially to be found much earlier in the literature,[1] Metzler's article is, to the best of my knowledge, the first instance of its being applied to the Keynesian system [2] The purpose of this note is first to make a minor refinement of Metzler's diagrammatic analysis, and then to point out that its geometric proof of stability in the small is actually inadequate—as is also the case for the corresponding proofs that accompany other phase diagrams in the literature

For convenience, let me carry out the discussion in terms of the version of the Metzlerian diagram that I have used elsewhere (Patinkin 1965, pp 232-33 et passim) The curve CC in figure 1 represents the locus of combinations of the rate of interest, r, and price level, p, which satisfy the equilibrium condition in the commodity market

$$F\left(Y_0, r, \frac{M_0}{p}\right) - Y_0 = 0, \tag{1}$$

I am indebted to William Baumol, Stanley Fischer, David Levhari, and Jerome Stein for comments on an earlier draft of this note Needless to say, none of them is to be held responsible for the views here expressed I wish also to thank the Central Research Fund of The Hebrew University and the Israel Academy of Sciences and Humanities for research grants to cover the costs of technical assistance

[1] See, e g , Marshall 1923, pp 341 and 353-55

[2] Though, of course, the first analytical treatment of the stability of the Keynesian system—and hence of the forces that lie behind the corresponding phase diagram—is to be found in Samuelson's famous 1941 article on the stability of equilibrium (pp 113-20)

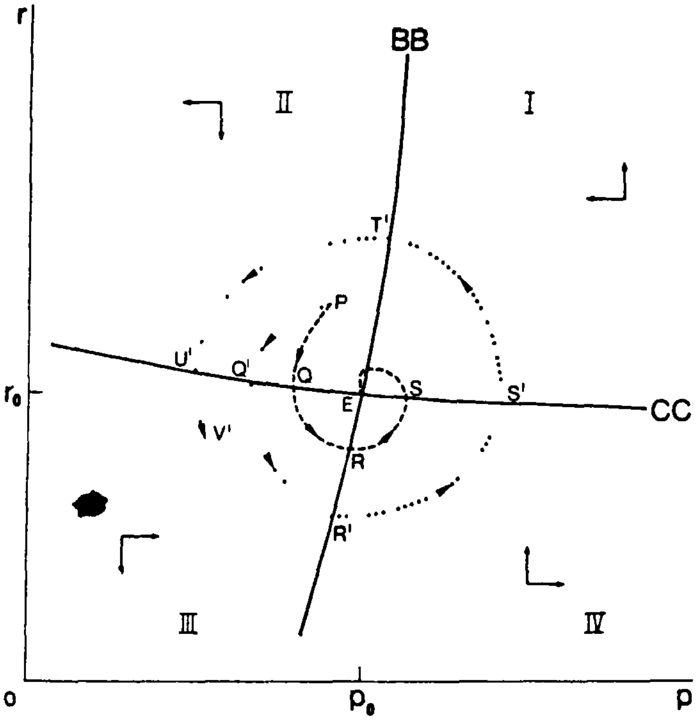


FIG 1

where Y_0 is the fixed full-employment level of output, and M_0 is the fixed quantity of money Similarly, BB represents the equilibrium condition in the bond market

$$B\left(Y_0, \frac{1}{r}, \frac{M_0}{p}\right) = 0 \qquad (2)$$

As usual, let the respective directions of the arrows in the four regions of the accompanying diagram represent the dynamic assumptions that whenever there is an excess demand (supply) in the commodity market, the price level will be driven upward (downward), and that whenever there is an excess demand (supply) in the bond market, the rate of interest will be driven downward (upward) Then, concludes Metzler, "When the economic system is out of balance at least one force is always operating to bring the variables of the system closer to the equilibrium point" (1951, p 105)

First, my refinement of this argument Immediately after the intersection of the dynamic adjustment path with one of the equilibrium curves, *both* forces must push in the direction of equilibrium For consider the dashed dynamic path that begins (say) at point P By assumption, as the path approaches CC, the extent of excess supply in the commodity market becomes less, indeed, when the path reaches the CC curve, this market is in equilibrium, so that the path must be vertical at the intersection point, Q Thus immediately after point Q, both p and r must move in the direction of the equilibrium point E A corresponding statement holds, *mutatis mutandis*, for the intersection point R, where the path must be horizontal

Furthermore, we can in a similar way see that the slope of the adjustment path throughout (say) region *III* must become increasingly flatter For as the path moves from intersection point Q, the extent of the excess demand in the commodity market increases, thus increasing the horizontal component of the dynamic vector, at the same time, the fact that the path is moving in the direction of BB means that the extent of the excess demand in the bond market is decreasing, thus decreasing the vertical component of the vector The consequent decrease (absolutely) in the slope of the adjustment path in region *III* will continue until the path becomes horizontal at its intersection with BB at point R, as already noted

The foregoing argument, even without this refinement, has generally been considered as constituting a proof of the stability of the system in the small—and I have certainly used it in this way throughout my own work But I have recently realized that this proof is not really valid—as is demonstrated (in fig 1) by the counterexample provided by the dotted dynamic path $PQ'R'S'$ that meets all of the preceding restrictions on its slope yet diverges from the equilibrium point E (This path has been drawn to take full advantage of the relative slopes of the CC and BB curves, which, as the reader may have guessed, have been chosen so as to correspond to the case of a divergent cobweb) Such a divergence is impossible in the analytical solution which, under the aforementioned dynamic assumptions, yields an adjustment path that must be contained within circles of ever decreasing radii, centered on the equilibrium point E The verticality of the adjustment path when it crosses CC—and its horizontality when it crosses BB—does imply that this must be the case in the respective neighborhoods of the intersection points Q', R', S', T',. But there does not seem to be any way of demonstrating graphically that this must continue to be the case as the path moves out of these neighborhoods In other words, the geometrical properties of the dynamic adjustment path as described above do not seem to be sufficient to prove that intersection point R' must lie within the circle centered on E and passing through Q'.[3]

What can, however, be graphically established is the divergence of the

---

[3] My colleague, Eytan Sheshinski, has suggested that this inability to provide a graphical proof of stability in the small may be due to the fact that the graphical argument does not make use of the assumption of linearity (in the neighborhood of the equilibrium point) on which the analytical proof of stability is based

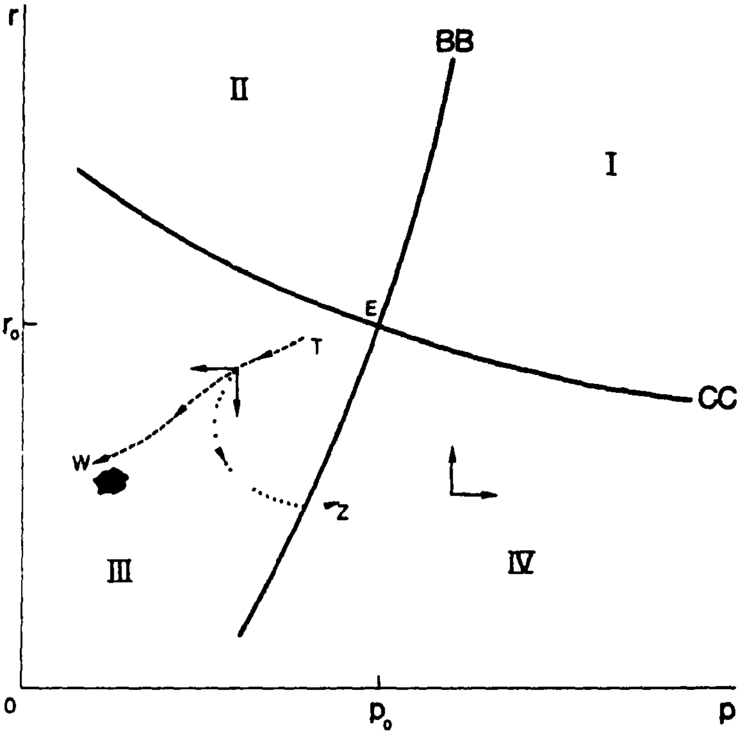


FIG 2

system in the case in which *both* dynamic forces push the variables away from their equilibrium values For example, if the dynamic path in sector *III* of figure 2 is subject to the forces represented by the arrows there, then the path must remain in this sector and continuously diverge from the equilibrium position In particular, it is impossible for the path to cross the *BB* curve and thus come under the convergent force of sector *IV* For, as we have seen above, in order to cross *BB*, the curve must, just before leaving sector *III*, move horizontally eastward (as on the path *TZ*), whereas the arrows show that the adjustment path in this sector (represented by *TW*) can never move in such a direction

References

Marshall, Alfred *Money, Credit, and Commerce* London Macmillan, 1923

Metzler, Lloyd A "Wealth, Saving, and the Rate of Interest " *J P E* 59 (April 1951) 93-116

Patinkin, Don. *Money, Interest and Prices* 2d ed New York Harper & Row, 1965

Samuelson, Paul A "The Stability of Equilibrium Comparative Statics and Dynamics " *Econometrica* 9 (April 1941) 97-120 Reprinted in *The Collected Scientific Papers of Paul A Samuelson* Vol 1 Cambridge, Mass M I T Press, 1966

Managerial Tenure under Private and Government Ownership in the Electric Power Industry

Louis De Alessi

George Washington University

Recent interest in the economic consequences of government regulation and government ownership of business enterprises has led to some fruitful applications of economic theory The effects of an active profit constraint on the behavior of a wealth-maximizing firm (Averch and Johnson 1962) have been subjected to increasingly rigorous theoretical scrutiny (e g , Takayama 1969) Though the discussion continues (e g , Zajac 1972), there is growing consensus about the general implications obtained (cf Baumol and Klevorick 1970). Concurrently, other researchers have investigated the choices of utility-maximizing decision makers within government-regulated, privately owned firms (Alchian and Kessel 1962), within government-owned business firms (Alchian 1965), and within regulatory agencies themselves (Eckert 1968) Empirical results to date (e g , Davies 1971, Peltzman 1971) support the conjecture that differences in the structure of property rights associated with these institutions provide government and private decision makers with different cost-reward structures and yield systematic differences in the outcome of the decision process In particular, the approach has been shown (De Alessi 1972) to be useful in explaining the published evidence relating to regulated and unregulated, private- and government-owned firms in the electric power industry (e g , Stigler and Friedland 1962, Moore 1970, Peltzman 1971)

One of the alleged implications of the utility-maximization hypothesis is that government decision makers enjoy longer tenure than their counterparts within privately owned firms (e g , De Alessi 1969, 1972) This paper presents a test of the hypothesis, using data drawn from the electric power industry Section I contains a brief derivation of the hypothesis,[1] Section II, a description of the data, and Section III, the test results Section IV contains the summary and conclusions

---

The research for this paper was supported by a grant from the National Science Foundation J L Gastwirth and an anonymous referee offered helpful suggestions

[1] This section is abstracted from a broader study of the economics of government control (De Alessi 1972)

I. The Hypothesis

An individual decision maker is hypothesized to maximize a single-valued, convex, twice-differentiable utility function The variables in the function reflect a variety of goals, such as health, security, prestige, and the welfare of others (De Alessi 1967) At the margin, all goods (sources of streams of utility to the chooser) are substitutable and, for each individual, at least some are scarce Fulfillment of the usual intra- and intertemporal equilibrium conditions implies that, among other things, the lower the cost of any good, the more will be acquired

Accordingly, a utility-maximizing manager has the incentive to increase the size and the duration of all job-related pecuniary and nonpecuniary streams of income For example, to increase his wealth, he has the incentive to increase the value of his marginal product by employing more staff and capital (Mayer 1960) and by increasing the size of the budget that he administers (Niskanen 1971) This strategy, in turn, is facilitated by reallocating activities from the future to the present (De Alessi 1969) and by increasing such variables as the output, the sales, and the size of the firm Similarly, a manager has the incentive to increase his nonpecuniary sources of income by hiring co-workers and employees with more desirable characteristics (as he sees it), by arranging more pleasant working conditions (as he sees it), and so on Up to a point, of course, such managerial behavior yields an increase in the firm's profits and is implied by the standard owners' wealth-maximization hypothesis The manager's ability to extend his activities beyond this point, thereby increasing his welfare at the expense of his employers' wealth, depends primarily upon the extent to which his contractual obligations (explicit and implicit) are enforced and upon the effectiveness of internal control forces

For present purposes, the crucial difference between private and government-owned (i e , political) firms lies in the relatively higher cost of transferring ownership shares in the latter (Alchian 1965) A taxpayer who wishes to change his portfolio of ownership in political firms normally must change the political jurisdictions in which he works and resides, or change the activities of the firms, or engage in similar high-cost activities As a result, property rights in political firms effectively are nontransferable This rules out specialization in their ownership and inhibits the capitalization of future consequences into current transfer prices, thereby reducing the incentive to detect and police managerial behavior which is inconsistent with employers' welfare Among other things, this implies that managers of political firms have greater opportunity to increase their own welfare at the expense of their employers' wealth, and that they find policies designed to increase noncapitalizeable returns to be particularly attractive [2] It might also be noted that control devices (e g , operating rules, line appropriations) internal to government firms become less market oriented as political considerations become more important in evaluating the outcome of the decision process

In the case of privately owned firms subject to government regulation, an active profit constraint implies that the gain to the owners from enforcing contracts falls Indeed, it is zero over some range Accordingly, the manager's transformation surface between pecuniary and nonpecuniary sources of utility—the boundary of the attainable region in his field of choice—shifts to the right, and he is able to increase all activities which contribute to his welfare Regulated firms, however, are more exposed than political firms to the discipline of the market, and their managers are more inhibited from making those choices which lower the wealth of the employers Among other things, the market for corporate control does act to constrain managerial behavior (e g , Hindley 1970) and to encourage market-oriented internal control devices (e g , De Alessi 1973) Moreover, the profit constraint need not be binding and, to the extent that it is, its effect is weakened by regulatory lag On the other hand, the managers of government firms, particularly those endowed with a polticallly influential clientele, can survive and even prosper in the presence of persistent deficits and significant economic losses Accordingly, they have greater opportunity for utility-maximizing behavior than the managers of private, regulated firms [3]

The salary and other contractual pecuniary rewards of government employees in general are subject to greater constraints than the pecuniary rewards of executives of private, regulated firms Among other things, the pecuniary rewards of government employees normally are subject to statutory ceilings, whereas those of the employees of private, regulated firms normally are not even reviewed by the regulatory agencies [4] One obvious consequence is that the opportunity cost of nonpecuniary sources of utility is lower within political firms More important, the manager of a political firm will find that the opportunity cost of activities designed to enhance the probability of his survival in office is also lower, and he will have greater opportunity to increase the present value of his pecuniary and nonpecuniary sources of income by increasing his job security For example, he will have greater incentive and opportunity to reward subordinates for their loyalty to him, to institutionalize decision making in order to dissipate responsibility for errors, and indeed, to institutionalize tenure (Alchian 1959) As a result, the manager of a political firm may be expected to survive longer in office

---

[2] Employers' incentive to enforce contracts may be expected to vary inversely with the political level (e g , local, state, federal) of the political organization in question

[3] The stream of benefits received by the manager of a political firm is less closely related to the future profitability of the firm, since he is unable to fully capitalize and capture job-related gains that will accrue in the future Accordingly, he attaches a lower value to such future gains relative to present gains than would otherwise be the case, and he has the incentive to increase the latter at the expense of the former Among other things, this implies that he will adopt lower discount rates in evaluating the firm's investment options (De Alessi 1969) Dams will be bigger and investments more capital consuming than would otherwise be the case Note that, for exactly the same reasons, employers also have less incentive to police such deviations

[4] Mr Jewell, director of public relations of the National Association of Regulatory Utility Commissioners, was queried After looking into the matter, he stated that the salaries of executives of regulated firms normally are not reviewed by regulators

II. The Data

The hypothesis that managers of government-owned firms have longer tenure than the managers of privately owned, government-regulated business firms was evaluated over the period from 1962 to 1971 using data drawn from the electric power industry

Time period —The choice of period was narrowed by the observation that prior to 1962 the Federal Power Commission did not collect the necessary information (e g , the names of managers) for government-owned electric utilities Although it might have been possible to obtain data for years prior to 1962 by sending questionnaires to the individual firms, it did not seem imperative to do so The sample period ended in 1971, the last year for which data were on file when this study began

The samples —Two random samples without replacement were chosen One sample consisted of 100 government-owned electric companies, mostly municipals, drawn from the 489 firms listed in the index of *Statistics of Publicly Owned Electric Utilities in the United States, 1962* (U S Federal Power Commission, 1963*b*) The population did not include federal utilities The other sample, also consisting of 100 firms, was drawn from the 223 firms listed in the index of *Statistics of Privately Owned Electric Utilities in the United States, 1962* (U S Federal Power Commission, 1963*a*)

Both samples excluded firms which did not report the desired information in 1962 (four government and four private firms), which ceased to report prior to 1971 (13 private firms), and which shared identical management with firms already in the sample (two government and 10 private firms)

Data —The basic data consisted of the name of the top executive of each firm for each year from 1962 to 1971 The initial intention had been to observe the tenure experience of individuals at several different levels of top management Government-owned electric companies, however, only report the name of their top executive (e g , superintendent, manager) to the Federal Power Commission As a result, comparisons between samples was limited to that slot The possibility of systematic differences in the tenure experience of individuals at different levels of

TABLE 1

FREQUENCY OF CHANGE IN TOP EXECUTIVE OF FIRMS IN THE ELECTRIC POWER INDUSTRY, 1962–71

OWNERSHIP CATEGORY	FREQUENCY OF CHANGE						SAMPLE SIZE
	0	1	2	3	4	5	
Political	36	42	17	3	1	1	100
Private	18	59	18	4	1	0	100

top management was explored within the privately owned electric firms using data on the top three executives (president and below) for the first 50 firms drawn

Length of tenure was indexed by the number of times that the name of the manager changed during the sample period A change was deemed to occur if the name reported in 1 year was different from that reported for the preceding year and if the name had not appeared before This last condition, designed to give individuals credit for coming back as well as for hanging on, had a negligible effect on the test results The frequency of changes is shown in table 1

III. The Evidence

Given the nature of the data (table 1), the χ^2 test seemed eminently suitable to evaluate the null hypothesis that the tenure of managers is the same for government and for privately owned electric companies [5] The test results ($\chi^2 = 8\ 89$, df $= 3$) indicate that the null hypothesis may be rejected at the 3 percent level of significance (two-tail).[6] Indeed, a comparison of the frequency distributions shown in table 1 suggests that managers of government-owned electric companies enjoy longer tenure, as hypothesized earlier

According to table 1, 36 percent of the government-owned firms did not change their top executive over the entire sample period, compared with 18 percent for privately-owned firms To examine the statistical significance of this difference, the data were combined in a 2 × 2

---

[5] *If degrees of freedom > 1, the χ^2 test may be used only if fewer than 20 percent of* the cells have an expected frequency of less than five and if no cell has an expected frequency of less than one The cells in the last two columns of table 1 fail to meet these conditions The prescribed remedy is to combine adjacent columns until the requirements are met (Siegel 1956, p 110), and this was done by consolidating the last three columns of table 1

[6] If a change is noted every time that the name of the manager is different from that reported for the preceding year, disregarding whether the name had appeared before, then the χ^2 test ($\chi^2 = 10\ 13$, df $= 3$) allows rejection of the null hypothesis at an even lower level of significance (2 percent, two-tail)

TABLE 2

FREQUENCY OF CHANGE IN TOP EXECUTIVE OF FIRMS IN THE ELECTRIC POWER INDUSTRY, 1962–71

Ownership Category	No Change	Change	Sample Size
Political	36	64	100
Private	18	82	100

contingency table (table 2) [7] An additional advantage of the 2 × 2 classification is that it allows an unequivocal examination of the directional hypothesis The χ^2 statistic, adjusted for continuity, was computed The test results ($\chi^2 = 7\ 33$, df = 1) are statistically significant at the 005 level (one-tail), and support the hypothesis that the managers of governm[illegible]t-owned electric companies enjoy longer tenure than the managers of privately owned electric companies

It is possible that the tenure experience of a top executive is systematically different from that of his immediate subordinates and that, if the experience of the latter were taken into account, the test results would be less favorable to the hypothesis Although there is no obvious theoretical justification for this argument, it seemed useful to examine its validity insofar as the data permitted Accordingly, the frequency of change of the top three executives of privately owned electric companies was collated over the sample period

A simple examination of the data (table 3) suggests that the tenure profiles are quite similar This is not surprising, since higher-level managers may be expected to move up together Be that as it may, the results are not inconsistent with the analysis presented earlier in this paper

Although the managers of government-owned electric firms enjoy longer tenure, it is at least arguable that this may be due to factors other than the ownership of the firm In the electric power industry, most of the political firms are municipal and—relative to the private firms—are generally smaller[8] and purchase larger amounts of the power they distribute [9] If smaller firms have a simpler hierarchical structure, their chief executives might be promoted to that position earlier in life and thus have a longer expected tenure Similarly, if firms which purchase

---

[7] The data shown in table 2 are the same for both definitions of change

[8] Using total assets and electricity produced (as defined below) as alternative measures of size, the *t*-test for the difference between two means was used to evaluate the null hypothesis that the average size of private firms is smaller than or equal to the average size of political firms The results ($t = 20\ 6$ for output and 7 3 for assets) allow rejection of this hypothesis at less than the 0 0005 level in the electric power industry, private firms generally are larger than political firms

[9] Of the firms sampled, one private (1 percent) and 48 political (48 percent) purchased all of the power they distributed

TABLE 3

FREQUENCY OF CHANGE IN THREE TOP EXECUTIVES OF PRIVATELY OWNED FIRMS IN THE ELECTRIC POWER INDUSTRY, 1962–71

RANK OF EXECUTIVE	FREQUENCY OF CHANGE						SAMPLE SIZE
	0	1	2	3	4	5	
Top	8	27	12	3	0	0	50
Second	6	30	8	5	0	1	50
Third	6	25	11	5	2	1	50

a larger portion of the power they distribute present their managers with simpler decisions and thus fewer possibilities of error, then their chief executives would also have a longer expected tenure

To examine the possible importance of these considerations, the tenure experience within each sample was related to the size of the firm and to the amount of electricity purchased Average tenure (T), defined as the length of the sample period divided by the number of chief executives over the period, was regressed upon the size of the firm, defined as the amount of electricity produced (P), and upon the amount of electricity received (R) Note that P and R together account for the total amount of electricity distributed by each firm [10]

$$T_{MUN} = 6\ 62469 - \underset{(1\ 04)}{0\ 02686\ P} - \underset{(0\ 98)}{0\ 07668\ R}, \qquad R^2 = 0\ 041 \quad (1)$$

$$T_{PVT} = 5\ 65197 - \underset{(1\ 12)}{0\ 00461\ P} + \underset{(0\ 43)}{0\ 00548\ R} \qquad R^2 = 0\ 013 \quad (2)$$

The results (the t-statistic is shown in parenthesis) are encouraging Although the size variable has the right sign, the level of significance is high—15 percent in the case of the political sample and 13 percent in the case of the private sample (one-tail test) [11] The variable for kilowatt-hours (kwh) received takes the opposite sign in the two regressions and is not statistically significant in either case Moreover, the coefficients of determination are not impressive

---

[10] The kwh produced and kwh received were measured by the average of the data reported in *Statistics* for 1962 and for 1970, the latter being the last year for which published data were available The data, rounded to the nearest 1,000, were divided by 100,000 for scaling

[11] As a further check, the regression was run using total assets (the average of the data reported in *Statistics* for 1962 and for 1970) as an alternative measure of size (A) Assets, rounded to the nearest 1,000, were also divided by 100,000 for scaling The results are practically identical to those obtained using output

$$T_{MUN} = 6\ 62594 - \underset{(0\ 96)}{0\ 29628\ A} - \underset{(0\ 88)}{0\ 07273\ R}, \qquad R^2 = 040 \quad (1A)$$

$$T_{PVT} = 5\ 64780 - \underset{(1\ 14)}{0\ 06412\ A} + \underset{(0\ 47)}{0\ 00609\ R} \qquad R^2 = 013 \quad (2A)$$

To explore matters further, the samples were pooled and the regression computed with a dummy variable for ownership (0 if private, 1 if political) added. As shown in equation (3), the ownership variable (D) has the predicted sign:

$$T = 5.74353 - \underset{(1.27)}{0.00572\,P} + \underset{(0.09)}{0.00131\,R} + \underset{(1.61)}{0.66686\,D}, \qquad (3)$$

$$R^2 = 0.040$$

and is the only one which is statistically significant at the 5 percent level or less The size variable has the predicted sign (managers of larger firms have shorter tenure), but the relationship observed could have occurred by chance with a probability of 10 percent The variable for kwh received also has the predicted sign, but it is not statistically different from zero The coefficient of determination again is relatively small Although additional tests would be useful,[12] these results are not inconsistent with the hypothesis that the managers of political firms have longer tenure

IV. Summary and Conclusions

Recent theoretical contributions have suggested that the utility maximization hypothesis may usefully be extended to predict the choices of decision makers within government-regulated and government-owned business firms Preliminary empirical results seem to support this approach Although the beginning is auspicious, the precise nature of the constraints on individual decision makers associated with alternative property rights structures and the range of application of the theory have yet to be fully established

According to the analysis presented in this paper, the utility-maximization hypothesis implies that managers of government-owned business firms have longer tenure in office than managers of regulated, privately owned firms The implication was tested using data drawn from the electric power industry over the period from 1962 to 1971 Although additional tests would be helpful, the evidence fails to deny it

This study supports present extensions of the utility-maximization hypothesis and provides additional evidence regarding the range of application of the theory and the nature of the constraints on individual decision makers

---

[12] For example, other specifications of the tenure-size relationship could be examined to provide additional assurance that size has been fully taken into account Better results might also be obtained by increasing the variation in tenure, perhaps via information prior to 1962, and by introducing additional variables Thus, some index of a manager's performance should be helpful Although the improvement expected did not seem large enough relative to the cost to justify more elaborate tests with these data, the points being raised would provide a useful base for further research

References

Alchian, A. A. "Private Property and the Relative Cost of Tenure." In *The Public Stake in Union Power,* edited by P. D. Bradley Charlottesville· Univ. Virginia Press, 1959.

—— "Some Economics of Property Rights." *Politico* 30 (December 1965) 816–29

Alchian, A A , and Kessel, R. A "Competition, Monopoly, and the Pursuit of Money " In *Aspects of Labor Economics* Special Conference Series, vol. 14, Princeton, N.J.· Princeton Univ Press (for Nat Bur Econ. Res), 1962

Averch, H., and Johnson, L L "Behavior of the Firm under Regulatory Constraint " *A E R* 52 (December 1962) 1052–1069

Baumol, W J , and Klevorick, A K "Input Choices and Rate-of-Return Regulation An Overview of the Discussion " *Bell J Econ and Management Sci* 1 (Autumn 1970) 162–90

Davies, D G "The Efficiency of Public versus Private Firms The Case of Australia's Two Airlines " *J Law and Econ* 14 (April 1971) 149–65

De Alessi, L "A Utility Analysis of Post-Disaster Cooperation " *Papers on Non-Market Decision-Making* 3 (Fall 1967) 85–90

——— "Implications of Property Rights for Government Investment Choices." *A.E R* 59 (March 1969) 13–24

——— "Government Ownership and Regulation Theory and the Evidence from the Electric Power Industry " Unpublished manuscript, 1972

——— "Private Property and Dispersion of Ownership in Large Corporations " *J Finance* 28 (September 1973)· 839–51

Eckert, R D "Regulatory Commission Behavior Taxi Franchising in Los Angeles and Other Cities " Ph D dissertation, University of California, Los Angeles, 1968

Hindley, B "Separation of Ownership and Control in the Modern Corporation " *J. Law and Econ* 13 (April 1970) 185–221

Mayer, T "The Distribution of Ability and Earnings " *Rev Econ and Statis.* 42 (May 1960) 189–95

Moore, T G "The Effectiveness of Regulation of Electric Utility Prices." *Southern Econ J* 36 (April 1970) 365–75

Niskanen, W A. *Bureaucracy and Representative Government* Chicago Aldine-Atherton, 1971

Peltzman, S "Pricing in Public and Private Enterprises Electric Utilities in the United States " *J Law and Econ* 14 (April 1971) 109–48

Siegel, S *Nonparametric Statistics for the Behavioral Sciences* New York McGraw-Hill, 1956

Stigler, G J , and Friedland, C "What Can the Regulators Regulate? The Case of Electricity " *J Law and Econ* 5 (October 1962) 1–16

Takayama, A "Behavior of the Firm Under Regulatory Constraint " *A.E R* 59 (June 1969) 255–60

U S , Federal Power Commission *Statistics of Privately Owned Electric Utilities in the United States, 1962.* Washington. Government Printing Office, 1963 (*a*)

———. *Statistics of Publicly Owned Electric Utilities in the United States, 1962* Washington. Government Printing Office, 1963 (*b*)

Zajac, E E "A Geometric Treatment of Averch-Johnson's Behavior of the Firm Model Reply." *A E R.* 62 (March 1972) 142

Utility Maximization Sufficient for Competitive Survival: Comment

Robert Piron

Oberlin College and University of California, Berkeley

In a recent issue of this *Journal*, Ladd (1969) demonstrates that utility maximization sufficiently guarantees competitive "survival," that is, it is not necessary to assume the profit maximization criterion. If he had proceeded further in following his logic, however, he would have shown that his model implies (*a*) that profit maximization also sufficiently guarantees competitive survival and (*b*) that dropping the profit maximization postulate, given his assumptions, changes nothing in the theoretical sense (though some interesting empirical issues may well be raised). These two observations stem from (*c*) the *theoretical identity* of profit maximization and utility maximization at the critical point in Ladd's model. Thus, his conclusion becomes trivial and does not constitute substantive criticism of traditional wisdom.

Using Scitovsky's classic construction (2), Ladd argues as follows (see fig. 1). Let $M_1(L, K_1) = M_1$ be the implicit form of the equation of the "minimum" indifference curve I (K_1 being the minimum acceptable level of satisfaction). Since the net income function for both the profit maximizer and the utility maximizer results from the maximization of M subject to the production function for various values of the parameter L, say L^*, we can write $M_\pi = M_\mu = M(L)$.

Thus (we ignore second-order conditions, Ladd does not), the necessary conditions for utility and profit maximization are, respectively,

$$dM(L)/dL = U_L/U_M \tag{1}$$

and

$$dM(L)/dL = dM_1(L, K_1)/dL \tag{2}$$

If (1) and (2) are satisfied for every value of $dM(L)/dL$ at the same value of L, the "Scitovsky condition" for double maximization for all L is

$$dM_1(L, K_1)/dL = -U_L/U_M, \tag{3}$$

and a firm satisfying (1) and $\pi = 0$ also satisfies (2) and $\pi = 0$, where "$\pi = 0$ means the firm is restricted to points on curve I: the zero-profit indifference curve" (p. 481). So far, so good. But, since $\pi = 0$ constrains the range of relevance to indifference curve I, the converse of the previous

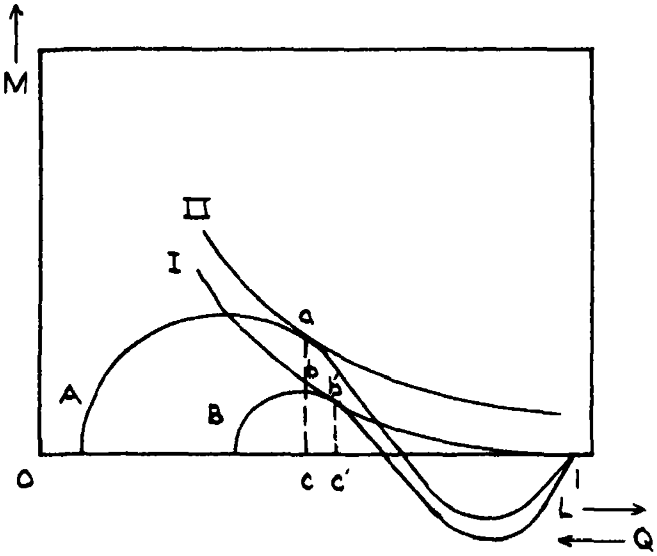


Fig. 1.—M measures entrepreneurial money income, L measures entrepreneurial leisure, Q measures the physical output of the firm.

result is also true: a firm satisfying (2) and $\pi = 0$ also satisfies (1) and $\pi = 0$. This is easily seen as follows. If the firm is maximizing profits at the level of zero, it must mean that we are at a point such as b'. Here, as contrasted with the case for profit maximization at some positive (or negative) level of economic profits, the tangents to the net income function and the indifference curve are coincident and not merely parallel. But, since the tangents are coincident, this means the indifference curve and the net income function must be tangent at b', which is precisely the condition for utility maximization, and, therefore, the two conditions become indistinguishable. Thus, by the constraining condition $\pi = 0$, utility maximization implies profit maximization, and, simultaneously, vice versa.[1] Thus, once the zero-profit constraint is effective,

---

[1] The equation

$$\frac{dM(L)}{dL} = -\frac{\partial U_1/\partial L}{\partial U_1/\partial M} \tag{4}$$

states that the slopes of the net income curve and the minimum indifference curve are equal to each other. As written, this is Ladd's equation (1.1), which was derived for the case of utility maximization. But this must imply

$$\frac{dM(L)}{dL} = dM_1(L, K_1)/dL = U_L/U_M \tag{5}$$

from the text above. Thus, when there is only one relevant indifference curve (the minimum curve I), the profit- and utility-maximizing conditions imply each other necessarily, a trivial case of the Scitovsky condition, which in general holds for all L.

the zero-profit equilibrium is formally identical with the utility-maximizing equilibrium by the rules of the Scitovsky model which dictate that "normal" profit be found as the vertical distance between the horizontal axis and the minimal indifference curve It seems that what Ladd has done, therefore, is to find a case of double maximization which need not necessarily imply the "Scitovsky case," though it is implied by it

Therefore, point b' in figure 1 (imagine indifference curve II and others above it as not being vertically displaced from curve I) would be a case of "double maximization" without the "Scitovsky condition" holding Does it matter if (3) does not hold? Not really, since regardless of the assumptions made concerning the argument of the utility function, long-run competitive equilibrium ($\pi = 0$) will occur at exactly that value of L which would hold if (3) were true Theoretically, therefore, it is as if the competitive firm were a profit maximizer (and a "double maximizer" at that) in the long run, and one suspects the triviality of Ladd's conclusion Finally, we need only appeal to the principle of Ockham's razor at this point and ask, What plausible criterion *would* a competitor maximize if not profit? Revenue, perhaps? Not if we want a finite result Market share? Impossible Security? The model does not define it in any meaningful terms Whatever the criterion, the model incorporating it must be more "complicated" than the profit maximization model, and since long-run results, as we have seen, are totally indistinguishable from one another, the choice of models seems clear on methodological grounds

References

Ladd, G W. "Utility Maximization Sufficient for Competitive Survival" *J P E* 77, no. 4 (July/August 1969) 478–82

Scitovsky, T "A Note on Profit Maximization and Its Implications" In *Readings in Price Theory* Homewood, Ill Irwin, 1952

On the Economics of a University: Toward the Determination of Marginal Cost of Teaching Services

Shlomo Brovender

University of Pittsburgh

I. Introduction

The current interest in the economic behavior of institutions of higher learning is motivated both by their financial plight and by the realization that economic theory is not well adapted to analyze the problems of nonprofit institutions The present publications in the economics of higher education tend to consider the university in the aggregate as a single, homogeneous plant The aim of this paper is to consider the university on the disaggregated level of departments and programs, and to suggest an alternative method of costing university teaching services (see also Brovender 1970)

II. The Distinction between Departments and Programs

Traditionally the university is viewed as a locus of departments, where each department pursues its own ends in undergraduate and graduate teaching, research, and other activities I assume that the primary outputs of the department are the number of credit hours it produces in a trimester or a year[1] and that the cost of unit of output can be measured in terms of average cost per credit hour [2]

In this paper I try to change the focus of attention from departments to programs For my purposes the program is defined by the final degree conferred on the student upon graduation, and conceptually many departments participate in each program (before only one department was usually associated with the program)

This paper is based on a study conducted at the University of Pittsburgh in 1970 I wish to thank Daniel B Rathbum, who as the director of Institutional Planning and Budgets in the University of Pittsburgh made this study possible, and Mark Perlman for his advice and support All errors and shortcomings are of course mine.

[1] I do not consider here the problems of output of research and community services but strictly the teaching activities

[2] If we consider the ratio of total credit hours to full-time equivalent faculty (FTE) to be a measure of average productivity, then average cost per credit hour = (salary bill/FTE)/(total credits/FTE) = average salary/average productivity

III. The Program Matrix

Assume that we can construct the following matrix, which I refer to as the A Matrix

The A Matrix is a rectangular matrix of coefficients in which the rows denote the academic program and the columns denote the academic departments The a_{ij} coefficients within the matrix are the average credit hours per period of time[3] required for a degree Hence, the matrix represents the curriculum structure of the university Consider table 1.

The relevant credit coefficients are those along the row vector (program vector) and not along the column vectors. A program major "buys" (uses) credits from different departments, but departments neither buy credits from nor "sell" (offer) credits to other departments This is a very important distinction that shows the validity of distinguishing between departments, which are both "production" entities and administrative entities, and "programs," which are functional entities When we say that the program major "buys" credits from other departments, essentially what we mean is that the program major imposes a load on other departments, and the extent of the load on each department will depend upon the number of students in the program, and the size of the coefficients a_{ij} within the matrix

In the A Matrix the sum of all rows

$$\sum_{j=1}^{n} a_{ij}$$

will be about the same magnitude, since the number of credits which a student is allowed to take in any period of time is constrained institutionally, that is, the student is not allowed to take more than a certain minimum and up to a certain maximum of credits—14–16 credit hours for a full load

The size of the matrix will depend on the number of identifiable

TABLE 1

A MATRIX

PROGRAM—DESTINATION	ACADEMIC DEPARTMENTS—SOURCE 1	2	N	
1	a_{11}	a_{12}	a_{1n}	$\sum_{j=1}^{n} a_{ij}$
2	a_{21}	a_{22}	a_{2n}	$\sum_{j=1}^{n} a_{ij}$
M	a_{m1}		a_{mn}	$\sum_{j=1}^{n} a_{ij}$

---

[3] The time dimension is important, since we can construct matrices for trimester, year, and 4 years, and by level of undergraduate/graduate instruction

programs and departments, but the numbers of departments and programs are not necessarily equal[4]

The distribution of the coefficients a_{ij} within the matrix is postulated as follows The largest coefficients will be distributed along the main diagonal ($i = j$) This is expected, since along the main diagonal the program major and the department which offers the majors coincide Off the main diagonal we expect to find an area of concentration of the coefficients[5] As we move farther away to the right or the left of the main diagonal, the coefficients will tend to be smaller and irregular For example, we can expect a student in premedicine to take English courses, but we hardly can expect a major in English to take premedicine courses Similarly, when we aggregate all the departments into larger units such as divisions or schools, we would expect to find a higher degree of concentration and interdependency within schools than between schools This results from the fact that some programs or schools tend to impose loads on the rest of the university while they do not offer, in a reciprocal way, credits to other majors from other programs So, the higher the degree of specialization of a school, the less the degree of interaction we can anticipate with the rest of the university

We expect that the coefficients a_{ij} within the matrix will tend to be relatively stable within the row vector over time To put it differently, we do not expect that the curriculum requirements will tend to change rapidly from period to period

Table 2 shows an actual A Matrix derived from a sample of 1,691 graduating students in the University of Pittsburgh in 1968–69

IV. Cost Considerations

Let $c_i = f(x)$ be a cost function, where c_i = salary bill in department i, and x_i = number of credits in department i Then $dc/dx = f'(x) = c'$ will be the marginal cost of producing 1 credit hour in department i, and $c_i/x_i = [f(x)]/x_i = \bar{c}$ will be the departmental average cost Furthermore, $x_i = x_{i1} + x_{i2} + x_{i3}$, where x_1 = lower-level undergraduate credits, x_2 = upper-level undergraduate credits, and x_3 = graduate credits[6]

---

[4] Since some departments are service departments, which do not have majors such as ROTC, continued education, etc

[5] The A Matrix refers mainly to undergraduate education In graduate programs we can expect a higher degree of specialization within departments and therefore less interdependence It is postulated that 90–95 percent of all coefficients will be along the main diagonal

[6] The credit hours used in the following regression equations were considered to be of equal weights The use of credit hours as a measure of output involves the following problems (1) determination of the absolute number of credits and primarily the distinction between X_2 and X_3, and (2) development of a nonarbitrary weighting scale Actually one of the purposes of the regression analysis is to develop a measure of relative costliness among X_1, X_2, and X_3 credits

TABLE 2

AN A MATRIX AVERAGE STUDENT CREDIT HOURS, UNDERGRADUATE PROGRAMS

	N	Human	Nat Sci	Soc Sci	Pre-med	Engin	Nurs	Educ	Pharm	Gen Stud	Inter-disc	ROTC	Total
Humanities	207	69 048	18 353	19 879		0 019	0 130	3 792		0 435	3 048	0 266	114 970
Natural sciences	360	34 753	58 492	18 933		0 478	0 242	2 619		0 597	2 703	0 833	119 655
Social sciences	409	34 888	20 597	51 472		0 178	0 037	3 147		1 675	3 169	1.108	119.271
Premedicine	26	30 346	43 000	16 269	5 538	1 000		2 154		0 115	3 577	0 231	102.230
Engineering	260	12 454	41 115	8 158		65 985	0 012	1 196		0 696	1 154	1 515	132.285
Nursing	60	19 950	14 800	5 450			64 267	1 123			0 750		106.340
Education	175	35 965	21 069	21 890		0 081	0 029	29 497		0 035	0 832	0 202	109.600
Pharmacy	35	16 171	19 143	8 886		0 229		0 800	102 686		1 457	0 571	149 943
General studies	161	25 969	24 242	23 000		0 957		2 242	0 012	4 689	0 242	0 075	81 428
Average	1,691	34 591	32 048	25 224	0 085	10 412	2 361	5 291	2 138	1 144	2 112	0 753	116 169

TABLE 3

RESULTS OF MULTIPLE REGRESSION ANALYSIS
TOTAL SALARIES = $b_0 + b_1$ (TOTAL UNDERGRADUATE) + b_2 (GRADUATE)

Group	b_0 (1)	Total Undergrad (2)	Graduate (3)	R^2 (4)
Humanities	249 4920 (77 8472)	4 5527 (1 6169)	74 8682 (25 9845)	78
Natural sciences	414 7910 (132 7140)	9 0277 (2 6964)	30 3118 (1 6174)	60
Social sciences	−42 8606 (109 6140)	18 9147 (2 6505)	17 2558 (15 1325)	78

NOTE —Numbers in parenthesis = standard errors

TABLE 4

RESULTS OF MULTIPLE REGRESSION ANALYSIS
TOTAL SALARIES = $b_0 + b_3$ (TOTAL CREDITS)

Group	b_0 (1)	Total Credits (2)	R^2 (3)
Humanities	373 9870 (66 3854)	8 1254 (0 8994)	73
Natural sciences	431 5740 (132 8490)	11 4350 (1 7621)	58
Social sciences	−42 8411 (107 1200)	18 7621 (2 1237)	78

NOTE —Numbers in parenthesis = standard errors

It should be emphasized that the above cost considerations follow the departmental definition of locus of activity and not the program definition

For exposition purposes only, tables 3 and 4 give the results of six regressions which attempted to estimate a cost function of the type described above in two variations

V. Synthesis

The existence of the A Matrix and both the cost coefficients from the regression equations (c') and the average cost coefficients ($\bar{c}$) will enable us to cost university programs on both an average-cost basis and a marginal-cost basis

For exposition purposes I will use a segment of 3 × 3 from table 2, with the corresponding coefficients from tables 3 and 4, to calculate the average program cost and "marginal" program cost[7] for a "typical" student in the divisions of humanities, natural sciences, and social sciences

---

[7] The coefficients of the regression equations are taken to represent the "marginal cost" of the division

TABLE 5

COMPOSITE TABLE

	No of Credit Hours				Cost per Credit Hour ($) MC* Coefficients of		
	Human (1)	Nat Sci (2)	Soc Sci (3)	Total (4)	ATC† (5)	Table 3 (6)	Table 4 (7)
Humanities	69 0	18 5	19 9	107 4	15 00	4 55 (1 62)	8 13 (0 90)
Natural sciences	34 7	58 5	18 9	112 1	18 87	9 03 (2 70)	11 44 (1 76)
Social sciences	34 9	20 6	51 5	107 0	17 87	18 91 (2 65)	18 76 (2 12)

SOURCE - Cols 1 4 from table 2, col 6 from table 3 col 2 col 7 from table 4, col 2
NOTE - Numbers in parentheses = standard errors
* Marginal cost
† Average cost per credit hour in the respective divisions

TABLE 6

CALCULATION OF PROGRAM COST

	No of Credits	Total Program Costs ($)			Program Cost per Credit Hour ($)		
		AC	MC_1	MC_2	AC	MC_1	MC_2
Humanities	107 4	1,736	855	1,143	16 19	7 97 (1 24)	10 66 (0 76)
Natural sciences	112 1	1,982	1,043	1,305	17 68	9 30 (1 56)	11 64 (1 02)
Social sciences	107 0	1,832	1,318	1,485	17 12	12 32 (1 47)	13 88 (1 12)

NOTE - AC = average cost, MC = marginal cost MC_1 refers to coefficients from table 3, and MC_2, to coefficients from table 4 Numbers in parenthesis = standard errors

The data in table 5 enable us to calculate the program costs in the humanities, natural sciences, and social sciences It should be realized that in essence the so-called program cost is nothing more than the weighted unit cost of the program, that is, the number of credits which were delivered by each "administrative" division (or department) times its unit costs for producing 1 credit hour [8]

Table 6 gives the results of these exercises The total program costs are equalized for the difference in the number of credits which are represented in the reduced matrix Now we are in a position to consolidate the arguments

---

[8] The program cost is equal to

$$\sum_{j=1}^{n} a_{ij}\bar{c}_{ij}$$

TABLE 7

PROGRAM COST (IN DOLLARS PER CREDIT HOUR)

	Department AC	Program AC	Program MC_1	Program MC_2
Humanities	15.00	16.19	7.97 (1.24)	10.66 (0.76)
Natural sciences	18.87	17.68	9.30 (1.56)	11.64 (1.02)
Social sciences	17.87	17.12	12.32 (1.47)	13.88 (1.12)

NOTE.—AC = average cost, MC = marginal cost. Numbers in parentheses = standard errors.

TABLE 8

DEVIATION OF MARGINAL COST FROM PROGRAM AVERAGE COST

	Program Average Cost	MC_1	MC_2
Humanities	100.0	49.2	65.8
Natural sciences	100.0	52.6	65.8
Social sciences	100.0	72.0	81.1

The first column in table 7 represents the customary way in which university services are priced, that is, the cost is the departmental cost (divisional cost) of producing 1 credit hour according to this customary calculation. The humanities courses are the cheapest, while the natural sciences, as some may agree, are the most expensive, and social sciences are intermediate.

The second column represents the program-weighted average costs. According to this calculation the order of costliness does not change. The *humanities programs are still the cheapest and natural sciences are the* most expensive, but the range of differences between each program is reduced substantially.

The last two columns in table 7 demonstrate the "marginal program cost" by utilizing the regression coefficients from tables 3 and 4. These calculations point immediately toward two important relationships. First and foremost, the marginal costs per credit hour are substantially below those of the department average cost and the programs' average cost. And second, we observe a reversal in the costliness of the program. The social sciences, rather than the natural sciences, have become the most expensive program.

Table 8 summarizes the extent of deviation of the marginal cost from the average program cost.[9]

---

[9] The fact that the marginal cost is smaller than the average cost may be indicative of the optimal level of output and the relative levels of utilization of capacity.

VI. Summary and Conclusions

The preceding sections have suggested that analyzing the cost of services of a university can be rewarding, particularly if modern techniques are used The conventional average costing method, which considers the department vectors as the relevant cost accounting unit, is not able to provide a good indication of the costliness of the teaching activities because of the following reasons (1) The average cost does not permit differentiation between the teaching and nonteaching activities that are carried out in the department (2) It is not able to differentiate between the levels of teaching (undergraduate or graduate instruction) (3) It does not take into account the random elements that contribute to the cost variability between departments (such as seniority structure, technology of teaching, department size, and market conditions which affect salaries) (4) Average-cost accounting identifies the department with the program and therefore neglects to consider the extent to which the interaction between departments and programs at any given moment affects the level of average cost

I think that the proposed approach overcomes most of the deficiencies of the average costing method primarily because it emphasizes the crucial role of the interaction between departments and programs in the university

Reference

Brovender, Shlomo "Interactions between Departments and Programs in Analyzing a University as an Economic Unit The University of Pittsburgh in the 1960's " Ph D dissertation, Univ Pittsburgh, 1970

Book Review

The Channels of Monetary Effects on Interest Rates By Phillip Cagan
New York Columbia University Press, 1972 $5 00

The stated purpose of this book is to examine, theoretically and empirically, two types of monetary effects on interest rates, the portfolio and the credit effect A third effect—the impact of inflation on nominal interest rates, or the Fisher effect—is also mentioned Most of the book is devoted to an analysis of the portfolio and credit effects Two additional chapters, which to me are less interesting and will not be discussed here, present a simple mathematical model of monetary effects on interest rates and an empirical investigation of the lags in monetary effects on interest rates and income

Under the portfolio effect, portfolios are rearranged in response to a rise in the rate of growth of money until interest rates have changed sufficiently to make the quantity of money demanded equal to the existing stock The changed interest rate affects income, which feeds back on desired money balances An increase in the money growth rate ultimately produces an increase in the rate of growth of nominal income, but the new equilibrium is at the same real rate of interest Under the credit effect, increased money growth is viewed as producing an increased supply of loanable funds The larger the supply of loanable funds, the lower is supposed to be the interest rate

The weakest part of Cagan's argument is his attempt in chapter 1 to make sense out of the credit effect The discussion relies on the concept of "loanable funds" Cagan writes "The concept of 'loanable funds' here is all-inclusive and is meant to be the flow of funds or credit (I use the terms interchangeably) appropriate to the determination of the general level of interest rates in the economy Only *net* purchases of financial claims by each person or business are included, such purchases can be financed by income receipts, net reductions in money balances, or the creation of new money" (p 3, n 2, emphasis in original) Simultaneous changes in assets and liabilities for an individual economic unit must be netted out, since such changes do not affect the excess demand for funds. Similarly, exchanges of one asset for another, or of one liability for another, must be netted out

There is, I believe, a serious problem with Cagan's argument First, note that only income receipts not spent on goods and services are available for net purchases of financial claims Now, consider the following example Miss X has a flow of nonspending of $1,000 per year which she invests in bonds issued by Mr Y, who is spending $1,000 per year more than his income receipts If the two should marry and become one economic unit while maintaining the same total of spending, has the flow of loanable funds declined by $1,000? If so, should we expect there to be an effect on market interest rates?

Marriages and mergers may or may not affect the aggregate of nonspending by the partners, but such events certainly do affect the size of the flow of loanable funds as defined by Cagan. Furthermore, any method of aggregation that nets out intrasector claims, as is done in the Federal Reserve's Flow of Funds Accounts, raises the same issue Finally, the problem cannot be avoided by taking the difference between aggregate income and aggregate spending on goods and services, this figure is simply the statistical discrepancy in the national income accounts[1]

The flow of loanable funds may either rise or fall as interest rates change depending on whether the demand or the supply function has shifted But Cagan never attempts to measure the total flow of loanable funds Instead, he works with that part of the flow accounted for by new money creation In chapter 2 he argues that whether the credit effect exists depends on the disposition of the revenue from new money creation The first part of this argument is an analysis of the revenue from money creation The revenue from government currency issues is examined, and then the revenue from deposit expansion under a variety of assumptions involving monopoly versus competitive banking Much of this discussion is designed to show that revenues (bank profits) from money creation are smaller than they appear at first blush because of the resources used in providing the services with which banks bid for deposits

The second part of chapter 2 consists of a discussion of the disposition of the revenue from money creation Cagan discusses the disposition of the government's revenue in terms of the extent of distribution effects between the government and the private sector and concludes that "no general conclusions can be drawn about the direction or magnitude of the effects " Similarly, the effects of private currency issues and deposit expansion also depend on whether distribution effects exist A rise in deposit expansion generates inflation, which transfers real resources from the holders of deposits to bank owners, a net effect on the supply of loanable funds can only occur if the gainers and losers react asymmetrically

Cagan's argument would, I believe, have been more powerful if he had begun his analysis with the very simple assumption of random distributions of government fiat money to the private sector This simple assumption has the advantage of letting the revenue from money creation accrue to the private sector The effect on the supply of loanable funds is then clearly a matter of distribution effects Cagan's elaborate analysis of commercial banks is relevant for the question of the *amount* of revenue but is largely irrelevant for the question of which way the distribution effects will go Furthermore, if the government raises revenue through money creation rather than through taxation, the problem is still one of distribution effects *within* the private sector because a tax on real money balances has been substituted for a tax on, say, income Distribution effects *between* the private sector and the government are relevant only if money creation is a substitute for government bond issues

The implications of the assumption of random money distributions have already been explored in the context of money growth models In terms of the exposition by Johnson (1969), under some assumptions a higher rate of money creation can accelerate real investment and produce a higher equilibrium capital/labor ratio accompanied by a lower real rate of interest The mechanism producing this result is that inflation lowers the demand for real balances, which in turn lowers the saving going into real balances and raises the saving going into capital formation The growth model mechanism is the exact opposite of that discussed by Cagan Cagan's concern is whether money creation adds to the supply of real loanable funds rather than whether it reduces the demand for real loanable funds by reducing the demand for real money balances However, both Cagan and Johnson stress that any long-run effects that do exist appear by virtue of the assumption that no interest is paid on money

The empirical work in chapter 3 is designed to show that monetary growth does indeed have a negative impact effect on interest rates The method used is to divide the reference cycle (the business cycle with turning points defined by the NBER) into nine stages Stage I is the trough, stage V the peak, and stage IX the next trough. (Stage IX is stage I of the next reference cycle.) The level of the interest rate and the monthly rate of change of money (M_2) are averaged for each stage. The change in the stage average for the interest rate is then regressed on the change in the stage average for money growth. Six different interest rate series were used; their correlations with money growth range from $-.16$ to $-.45$.

In chapter 4 Cagan attempts to test for the credit effect. He constructs a consolidated monetary system balance sheet by consolidating the accounts of the Federal Reserve, the Treasury, and the commercial banks. The fully consolidated balance sheet has three principal assets and four principal liabilities. The assets are (*a*) gold (plus special drawing rights and foreign exchange at the present time), (*b*) commercial bank loans and investments other than Treasury securities, and (*c*) physical assets owned by the Treasury, Federal Reserve, and commercial banks. The liabilities are (*d*) deposits held by the public in commercial banks, (*e*) Federal Reserve and Treasury currency outside the monetary system, (*f*) Treasury securities outside the monetary system, and (*g*) the net worth of the Treasury, the Federal Reserve, and the commercial banks. Credit granted by the consolidated monetary system to the remainder of the economy is defined to equal item *b* less item *f*. Call this variable *E*, and its average monthly change over a reference cycle stage, *dE*.

Cagan runs regressions of the form $\Delta i = a_0 + a_1\Delta(dE/M) + a_2\Delta(dM/M)$, where Δ refers to the difference between the stage averages. The a_1 coefficient is supposed to represent the credit effect. Using the sample period 1919–66, Cagan finds that the credit effect is 10–40 percent of the monetary effect depending on the interest rate series used (i.e., the a_1 coefficient is 10–40 percent of the a_2 coefficient). Both a_1 and a_2 are negative, as expected.

This argument seems to me to confuse two meanings of "credit effect." Cagan's argument in chapters 1 and 2 has nothing to do with whether money creation is associated with an increase in the earning assets of the consolidated monetary system. There is no change in *E* if new money is distributed randomly from airplanes or systematically through tax cuts. The argument underlying "credit effect *A*" is that money creation will depress the real rate of interest by producing distribution effects ("forced saving") whereas under the portfolio effect the real rate would eventually return to its initial level with a new higher equilibrium inflation rate. Cagan's test is relevant to "credit effect *B*," the argument that it is credit creation rather than money creation that drives economic activity. The extreme form of this argument is that money creation per se does nothing, while money creation through government deficits, or through new bank lending, is expansionary.

The argument for credit effect *B* is in fact very close to the Keynesian argument that autonomous expenditures drive economic activity. An exogeneous rise in *E* can occur without a change in the money stock (the sum of balance sheet items *d* and *e*) if, for example, the government reduces its expenditures so that government debt is retired and net worth rises. The decline in autonomous expenditures that increases *E* depresses interest rates according to both the Keynesian model and Cagan's regressions.[1] With the exception of international capital flows, there is no way for *E* to change over time with *M* fixed without there being an associated change in private or government expenditures (or taxes). In a general equilibrium model with three goods—commodities, money, and securities—analysis may be conducted in terms of any two of the three markets. The argument for credit effect *B* is simply the same as the argument that both monetary and fiscal policy affect interest rates.

---

[1] Whether income changes depends on the extent to which the interest rate change crowds out (or in) other expenditures.

Another version of credit effect *B* is that bank credit plays a special role. Cagan tested two versions of this theory by defining a variable E' to equal total commercial bank loans and investments plus the Federal Reserve's securities portfolio, and by running a multiple regression splitting E' into commercial bank and Federal Reserve components. In both versions the credit effects are estimated to be somewhat larger than before. I find these results unconvincing. Bank credit can change relative to the money stock only as a result of changes in the currency/deposit and reserve/deposit ratios.[2] Both these ratios are probably related to interest rates, and such relationships seem more plausible explanations than "credit effects."

It appears, then, that Cagan has not in fact tested credit effect *A*. However, it is not clear that it is possible to distinguish credit effect *A* from the portfolio effect. Since both effects are properly defined in terms of the real rate of interest, a test for a long-run credit effect *A* is inevitably entangled in testing for the Fisher effect. In principle, distinguishing between credit effect *A* and the portfolio effect in the short-run might depend on distinguishing monetary effects on the level from those on the rate of change of the interest rate. Under the portfolio effect the interest rate ought to change so long as there is monetary disequilibrium; if (equilibri[illegible] money demand differs from money supply primarily because of changes in the latter, then it might be reasonable to argue that the portfolio effect should show up as a negative relationship between the rate of change of the interest rate and the rate of change of the money stock. Credit effect *A*, on the other hand, should appear as a relationship between the level of the interest rate and the rate of change of money. Unfortunately, all of the tests reported in this book relate the level of the interest rate to the rate of growth of money, and so no comparison along the above line is possible.

What have we learned from all this? The credit-effect arguments make sense to me only if I substantially restate Cagan's argument, and I am not at all sure that my restatement reflects what he had in mind. I cannot think of any purpose for which I would use the estimated coefficients in the regressions of interest rates on money growth. The primary value of Cagan's empirical work would seem to be that there can be little remaining doubt about the major direction of causality between the money stock and the interest rate. From what we know about the operation of the U.S. monetary system, if the money stock were an endogenous variable, then surely the short-run correlation between money and interest should be positive. Cagan has examined this issue in almost every conceivable way and has found that the short-run correlation between the interest rate and money growth is negative. This finding may not have been Cagan's aim, but monetarists should be happy to have this one securely salted away.

WILLIAM POOLE

Federal Reserve Bank of Boston

Reference

Johnson, H. G. "Money in a Neo-classical One-Sector Growth Model." In *Essays in Monetary Economics*. Cambridge, Mass.: Harvard Univ. Press, 1969.

---

[2] With the growth in nondeposit liabilities in recent years we might include such liabilities as a separate item in the consolidated balance sheet of the monetary system. In constructing the variable *E* this item would be subtracted out; *E* would be unaffected by equal growth in bank loans and bank nondeposit liabilities. However, nondeposit liabilities might not be subtracted out in constructing E'. There would not appear to be a good reason for this treatment except insofar as improvements in financial markets, including intermediation, reduce the demand for money. Given the money stock, a change in bank credit is more likely a symptom than a cause.

An Econometric Evaluation of Sulfur Taxes

James M Griffin

University of Pennsylvania and University of Houston

In contrast to the standard comparative static analysis of alternative policy instruments for dealing with air pollution, the present study provides a dynamic framework of analysis through the use of an econometric model An econometric model of the electric utility industry is utilized to calculate the welfare gains associated with the imposition of a sulfur tax as enumerated in the Pure Air Tax Act of 1972 The dynamic properties of a recent proposal by Baumol and Oates which would combine an ambient air standard and a charges policy are also examined

I. Introduction

There is considerable doubt as to what the appropriate policy tools for dealing with sulfur oxide emissions are Recent proposed legislation such as the Pure Air Tax Act of 1972 calls for a fixed tax rate [1] Alternatively, several economists, including Baumol and Oates (1971), propose the use of both an ambient air standard and a tax which the pollution control agency would adjust periodically to achieve target reductions in pollution. At least four major problem areas constrain meaningful comparisons of these alternative policy instruments Applied welfare analysis is circumscribed both by the lack of reliable estimates of the damage function and by deficiencies in the estimation of control costs Furthermore, Ayres and Kneese (1969) suggest that second-best considerations are likely to be

---

I wish to express thanks to F G Adams, George Daly, R William Thomas, Tom Mayor, and Henry Steele for numerous helpful comments This project was partially financed by a Faculty Development Research Grant from the University of Houston and by a grant from the Energy Policy Project of the Ford Foundation

[1] Specifically, this proposal calls for a 15¢/pound tax on sulfur emissions beginning January 1, 1976, for those areas that do not meet the primary standards and a 10¢/pound tax effective in those areas that meet the primary standards but not the secondary standards as set forth in the Clean Air Act (see *Congressional Record* 1972)

important, implying an inherent weakness of the traditional partial-equilibrium approach to welfare analysis. Finally, while the theoretical literature examines alternative policy instruments, it does so within a comparative static framework, thereby overlooking the inherently dynamic nature of the pollution control problem.

This paper seeks to demonstrate how the overly restrictive assumptions of comparative static analysis can be remedied by an econometric model As well as providing a dynamic framework, an econometric modeling approach provides a method of estimating control costs Specifically, simulations with an econometric model of the electric utility industry (Griffin 1974*a*) are presented to assess the likely welfare effects of the Pure Air Tax Act of 1972 and the dynamic properties of the Baumol and Oates approach Study of the electric utility industry is particularly appropriate because in 1970 power plants contributed 54 percent of the nation's su[illegible] dioxide emissions [2]

Theoretically, one should be able to recommend a tax on pollutant emissions which would minimize the sum of the damage costs, the polluters' control costs, and the regulatory agency's administrative costs Unfortunately, our understanding of these cost relationships is generally not sufficient to allow such a calculation Considerable effort has been devoted to determining the damage function of sulfur dioxide and particulate emissions through analysis of property-value data These studies generally obtain the expected negative relation between property values and pollution levels,[3] but the estimated coefficients are highly sensitive to the specification of the model There is also some question as to the shape of the marginal damage function Ridker and Henning (1967) found the function to be flat, while Anderson and Crocker indicate that it declines over low and intermediate ranges of pollution (Crocker 1971, p 70) While recognizing its importance, this paper does not attempt to add to the damage function literature but, rather, draws upon existing studies when necessary Fortunately, the policy conclusions of this study hold for a wide variety of assumptions about damage functions

In contrast to the attention given the estimation of damage costs, the problems of estimating control costs have been generally overlooked Perhaps because engineering data are a necessary input, the estimation of control costs has tended to be perceived as strictly an engineering question. For example, in the case of a sulfur tax, the estimation of control costs has been viewed merely as the cost of reducing a fuel of high-sulfur content to low-sulfur content after some specific period of transition [4] Yet economic theory would suggest the possibilities of factor input substitution and final product substitution by the consumer, which effects would tend to reduce control costs.

---

[2] For an informative review, see U.S. Congress (1970, pp. 172–88)

[3] See Ridker (1967), Ridker and Henning (1967), Anderson and Crocker (1970), Wieand (1970), and Crocker (1971).

[4] For an example, see Hausgaard (1970).

In addition to affecting the estimation of control costs, the presence of various lagged substitution effects points to weaknesses in the traditional comparative static analysis. If pollution charges are to be adjusted annually to achieve some ambient air standard, knowledge of the dynamic as well as long-run multipliers associated with the various substitution effects is required. From experiences with stabilization policy, it is conceivable that certain policy instruments may have destabilizing effects because such instruments are sensitive to unknown and variable dynamic phenomena Since the econometric model used in this study attempts to characterize both current and lagged-period substitution effects, it is appropriate to consider the dynamic implications of proposals to vary taxes in order to achieve stated emission levels

Section II presents a simplified description of the basic structural relationship present in the econometric model as well as the assumptions and estimation techniques underlying it In Section III, the probable welfare and distributional effects of the recently proposed sulfur tax are estimated for alternative technological assumptions and tax rates. The dynamic properties of the Baumol and Oates approach are considered in Section IV as taxes are adjusted annually to meet some specific sulfur emission constraint In Section V, factors omitted from this study, such as regional and seasonal effects and second-best considerations, are discussed A brief summary is given in Section VI

II. The Model

The simulation results reported herein are based on an econometric model of the electric utility industry (Griffin 1974*a*) and subsequently modified in Griffin (1974*b*) to consider the sulfur tax issue. The interested reader is referred to the above two sources for a detailed description of the model The description here is intended to acquaint the reader with the basic characteristics of the model and to illustrate the major substitution possibilities, which will become important in Section IV.

After modification for sulfur tax analysis in Griffin (1974*b*), the model consists of 33 equations of which 12 are stochastic. Of the 27 exogenous variables, the four major macroeconomic variables are linked to the *Wharton Annual and Industry Forecasting Model* (Preston 1972), which provides simulation inputs to 1981 With the exception of one equation, which follows MacAvoy (1969) in using a cross-section time series panel, the sample consists of annual observations for the period 1950–71 To distinguish short-run and long-run elasticities, Almon polynomial distributed lags are estimated in several equations using maximization of the coefficient of determination as the basis for the choice of lag structure

Two-stage least squares is the principal estimation approach because of the numerous simultaneous relationships in the model. A 5 percent t-test criterion is used for inclusion of variables that are first determined as relevant on theoretical grounds.

In applying a model to long-run simulations for 15 years beyond the sample period, a number of pitfalls are encountered. First, the choice of values for exogenous variables is subject to considerable error, and unfortunately the policy implications are not independent of this choice. A second source of concern is that the estimated structure based on a restricted range of observations will prove deficient over an extended sample range. As shown in Griffin (1974*a*), the simulation period utilizes ranges of exogenous variables well beyond those in the sample period. This is especially true of the exogenous values for the respective fossil fuel prices. Unfortunately, little can be done except to warn the reader of the potential gravity of this problem and to label the results as rough approximations.[5]

The electricity model contains endogenous determination of the demands for electrical power, the choice of the primary fuel input sources to meet these generation requirements, and determination of electricity prices. In the model, prices of fuel and capital inputs, which are exogenous, play an important role. Factor input prices not only affect the choice of fuel type and the efficiency of conversion to electricity, but indirectly affect the price of electricity and consequently, the demand for electricity. Sulfur tax analysis is achieved by adjusting the various fossil fuel prices for a given tax rate and exogenous set of technological control cost assumptions.

To acquaint the reader with the model and yet avoid the detail of Griffin (1974*a*, 1974*b*), the model is condensed here from 33 equations to 17 and presented in general functional form. This condensation implies (1) that numerous less important endogenous variables appear as explanatory variables but do not reappear as dependent variables, (2) that only major identities are included, and (3) that some behavioral equations appear in forms other than those in which they were estimated. Exogenous variables are indicated by an asterisk following the variable. The model is best characterized as consisting of the following four blocks: electricity demand, fuel source for generation, electricity price determination, and welfare calculation for sulfur taxes. No attempt is made to endogenize the demand for factor inputs other than fuel. Consequently, the simulations presume that capacity will grow at a rate sufficient to meet generation requirements.

---

[5] Chow tests for stability of coefficients over the sample period were performed, but they provide only weak evidence for observations over much different sample ranges.

The demand for electricity is separated into two distinct functional forms. In equation (1), the residential demand for electricity, *DR*, depends on distributed lags, $\beta(L)$ and $\gamma(L)$,[6] involving per capita real disposable income, YD^*/N^*, and the average residential price of electricity, *PR*, deflated by the GNP deflator, P^*, the residential stock of air conditioners, *KR*, and population, N^*

$$DR = f\left[\beta(L)\frac{YD^*}{N^*}, \gamma(L)\frac{PR}{P^*}, KR, N^*\right] \tag{1}$$

The residual, consisting primarily of industrial and commercial demands for electricity, *DI*, is a function of the level of economic activity as measured by real GNP, *X*, the stock of commercial air conditioning, *KC*,[7] and a distributed lag on the average industrial and commercial price of electricity, *PI*, deflated by the GNP deflator, P^*

$$DI = f\left[X^*, KC, \beta(L)\frac{PI}{P^*}\right] \tag{2}$$

Given a determination of residential and industrial demands for electricity, we can determine total generation requirements to be met by electric utilities, *TG*, from the following relationship by specifying net imports, $Z1^*$, transmission losses, *Z2*, and own-use generation by industries, $Z3^*$

$$TG = DR + DI - Z1^* + Z2 - Z3^* \tag{3}$$

It also follows in equation (4) that total generation by electric utilities, *TG*, must be met from one of the following five sources nuclear generation, *NG*, natural-gas-fired generation, *GG*, oil-fired generation, *OG*; coal-fired generation, *CG*, and hydroelectric generation, HG^*

$$TG = NG + GG + OG + CG + HG^* \tag{4}$$

In equation (5), nuclear generation, *NG*, depends on the previous level of installed nuclear capacity, which is reflected in the previous year's nuclear generation, NG_{-1}, and the increment in generation requirements to be met by nuclear and fossil fuels, $\Delta(TG - HG^*)$, multiplied by the share of new nuclear capacity relative to nuclear and fossil capacity additions, $\Delta CN/(\Delta CN + \Delta C)$

---

[6] $\beta(L)$ is the lag operator indicating a distributed lag on *X* as follows· $\beta(0)X + \beta(1)X_{-1} + \beta(2)X_{-2}$ $\beta(n)X_{t-n}$ Note that the subscripts below *X* indicate the number of periods lagged

[7] Both residential and commercial stocks of air conditioning are endogenous and depend on lagged income and GNP variables, respectively Their direct inclusion in (1) and (2) allows the adjustment of the user cost of fossil capital for peak shaving in (6).

$$NG = NG_{-1} + \Delta(TG - HG^*) \times \frac{\Delta CN}{\Delta CN + \Delta C}\left[\left(\frac{PNF^*}{PFF \cdot HR}\right)_{-7}, \left(\frac{PNK^*}{PFK^*}\right)_{-7}\right] \quad (5)$$

The share of new nuclear capacity is assumed to depend on the relative prices of nuclear fuel, PNF^*, relative to fossil fuel, PFF HR [8] and the relative capital prices of nuclear, PNK^*, and fossil, PFK^* The subscripts indicate that new capacity expansions depend on relative prices 7 years previously

Before explaining the generation rate by the various fossil fuels, it is instructive to determine the efficiency with which Btu inputs of fossil fuel are converted into kilowatt-hours of electricity This efficiency measure is characterized by the heat rate, HR, which indicates the number of Btu's of fuel input needed to generate a kilowatt-hour of output The heat rate, HR, depends on distributed lags involving the rate of new fossil capacity additions, $\Delta C/C^*$, and the user cost of capital in fossil plants adjusted for peak-shaving requirements, UC, relative to the average price of fossil fuels, PFF

$$HR = f\left[\beta(L)\frac{\Delta C^*}{C}, \gamma(L)\frac{UC}{PFF}\right] \quad (6)$$

Given the exogenous value for hydroelectric generation, HG^*, the rate of nuclear generation, NG, in (5), and the determination of total generation, TG, in (3), we now need to determine oil, coal, and gas generation in (4) Natural-gas generation, GG, is assumed to depend on the conversion efficiency, HR, and on supply conditions of "low-priority" gas for boiler fuel uses, AG^* ·

$$GG = f(HR, AG^*) \quad (7)$$

In equation (8), coal generation, CG, depends on the conversion efficiency of fossil fuels, HR, the remaining market to be met by coal or fuel oil, $OG + CG$, a dummy sulfur pollution control variable, DU^*, for the period 1969–71, and a distributed lag on the price differential between fuel oil and coal, $PO^* - PC^*$

$$CG = f[HR, OG + CG, DU^*, \beta(L)(PO^* - PC^*)] \quad (8)$$

Given HG^*, NG, GG, CG, and TG, oil generation, OG, is treated as a mathematical residual and solved for in equation (4) The assumption that oil is the marginal fuel source certainly is not inconsistent with recent experience.

---

[8] PFF is the average price per million Btu's which must be adjusted by the heat rate HR, a measure of conversion efficiency, in order to provide a fossil fuel cost in kilowatt hours

The electricity price block of the model relates the fuel choices and the exogenous values for the prices of the various fuels to the price of electricity. Equation (9) expresses the simple identity that the average price of fossil fuel in Btu's, *PFF*, which appears in (5) and (6), depends on the prices of natural gas, PG^*, fuel oil, PO^*, and coal, PC^*, weighted appropriately:

$$PFF = PG^* \frac{GG}{GG + OG + CG} + PO^* \frac{OG}{GG + OG + CG} + PC^* \frac{CG}{GG + OG + CG} \tag{9}$$

Similarly, in equation (10), we calculate the average price of all fuel inputs on a kilowatt-hour output basis, *PF*, by weighting appropriately the price of nuclear fuel, PNF^*, and the price of fossil fuel adjusted to an output basis, $PFF \cdot HR$.

$$PF = PNF^* \frac{NG}{TG} + (PFF\ HR) \frac{GG + OG + CG}{TG} \tag{10}$$

The average residential price of electricity, *PR*, is treated as a markup over these unit fuel costs, *PF*, capital costs, *AUC*,[9] and a variable to reflect the declining rate schedule as per customer purchases increase, DR/N^*

$$PR = f\left(PF, AUC, \frac{DR}{N^*}\right) \tag{11}$$

Similarly, the average industrial price is treated as a markup over costs with a scale proxy, DI/X^*

$$PI = f\left(PF, AUC, \frac{DI}{X^*}\right) \tag{12}$$

The welfare block of the model calculates the net welfare gains arising from a sulfur tax In equation (13), control costs are calculated as a function of the use of coal and oil inputs, *OG*, *CG*, and *HR*, and the exogenous control costs per 1 million Btu's for oil and coal, CCO^* and CCC^*

$$CC = f(CCO^*, OG, CCC^*, CG, HR) \tag{13}$$

As illustrated later, the welfare losses, *WL*, from a tax consist of the control costs, *CC*, plus the loss in consumer surplus, $f(\ \ )$, which depends on price and quantity differentials from a base case assuming the air to be a free good

---

[9] Based on a weighted average of past values of the user cost of capital for utilities as provided by the annual Wharton model (Preston 1972)

$$WL = CC + f(PI - \widehat{PI}^*, \widehat{DI}^* - DI, PR - \widehat{PR}^*, \widehat{DR}^* - DR, P^*) \tag{14}$$

The GNP deflator P^* is included to provide welfare calculations in constant 1971 dollars

In order to calculate the welfare gains from imposing a sulfur tax, it is first necessary to calculate the emissions of sulfur from a given tax Merely by using coal and oil generation, OG and CG, together with the heat rate, HR, the Btu inputs of the two sulfur-bearing fuels is determined Sulfur oxide emissions, SU, can then be calculated given the fuel inputs and the percentage sulfur in each fuel remaining after control processes, $\%SC^*$ and $\%SO^*$, as follows

$$SU = f(\%SC^*, CG, \%SO^*, OG, HR) \tag{15}$$

Welfare gai[illegible] WG, resulting from the improvement in air quality simply equals the incremental damage cost of a ton of sulfur, IDC^*, multiplied by the reduction in sulfur emissions as a result of the tax, $(\widehat{SU}^* - SU)$

$$WG = IDC^*(\widehat{SU}^* - SU) \tag{16}$$

Net welfare gains, NWG, is simply the difference between welfare gains, WG, and welfare losses, WL

$$NWG = WG - WL \tag{17}$$

Sulfur tax simulations are easily performed because the prices of all fuels are exogenous to the model For a given tax rate, the price of fuel oil and coal are adjusted to reflect their respective control costs, CCO^* and CCC^*, as well as the tax paid on the residual sulfur content, which involves the tax rate applied to the percentage remaining sulfur, $\%SC$ and $\%SO$ Note that the assumption of exogenous fuel prices, independent of the sulfur tax rate, is somewhat unrealistic, but to endogenize such elements is well beyond the scope of this paper

The model includes many of the substitution effects discussed in the introduction that depend on relative factor and product prices For example, in addition to the substitution of low-sulfur fuel for a high-sulfur fuel, a sulfur tax will affect the price differential between coal and fuel oil, causing an interfuel substitution effect in (8) Similarly, a sulfur tax will raise the price of fossil fuel, PFF, and induce a substitution for nuclear generation in (5) The model also includes the possibility of substitution between fuel inputs and capital inputs, as the heat rate in (6) depends on the relative prices of capital and fossil fuel The lagged price variables in both the residential and industrial demand equations imply that consumers will respond to the higher electricity prices by substituting

TABLE 1

SELECTED SHORT- AND LONG-RUN ELASTICITY ESTIMATES (CALCULATED ABOUT SAMPLE MEANS)

Variables	Short-Run Elasticity	Long-Run Elasticity
$DR \mid PR/P$	-0.06	-0.52
$DI \mid PI/P$	-0.04	-0.51
$\frac{\Delta CN}{\Delta CN + \Delta C} \mid PNF/PFF\ HR$	0	-1.40
$\frac{\Delta CN}{\Delta CN + \Delta C} \mid PNK/PFK$	0	-0.86
$HR \mid UC/PFF$	0.03	0.30
$CG \mid PC$	-0.02	-0.20
$CG \mid PO$	0.02	0.24

other goods for electricity-intensive goods. The treatment of substitution effects is incomplete, since interfuel substitution among coal, fuel oil, nuclear, and the other power sources—natural gas and hydroelectricity—does not occur in the model. This omission is not serious, however, as supply conditions are likely to restrict their substitutability.

While the model includes the primary substitution effects, it is important to realize that these substitution effects do not occur instantaneously. To illustrate this, table 1 reports the elasticities of various variables in the model with respect to given factor or product price changes for 1 year (a short-run elasticity) and after all adjustments (a long-run elasticity).[10] The estimates are calculated about sample means. For example, the first entry shows the elasticity of residential electricity demand, *DR*, with respect to the real price of electricity, *PR/P*. These estimates indicate that substantial substitution effects are present in the long run but, in the short run, choices as to fuel consumption and electricity demand are highly circumscribed by existing capital configurations.

III. The President's Sulfur Tax

In this section, we consider the welfare and distributional effects of the Pure Air Tax Act of 1972 as they relate to electric utility generation. The model is simulated over the period 1972 and 1985 with varying technological assumptions about sulfur control costs and with alternative sulfur tax rates, then these results are compared with a base case solution in which the air is treated as a free good. The relative difference between the alternative simulations are probably much more informative from a policy perspective than the absolute welfare losses and gains, which are conditional on quite restrictive assumptions.

[10] For a comparison of these estimates with other studies such as by Anderson (1972) and Wilson (1971), see Griffin (1974a).

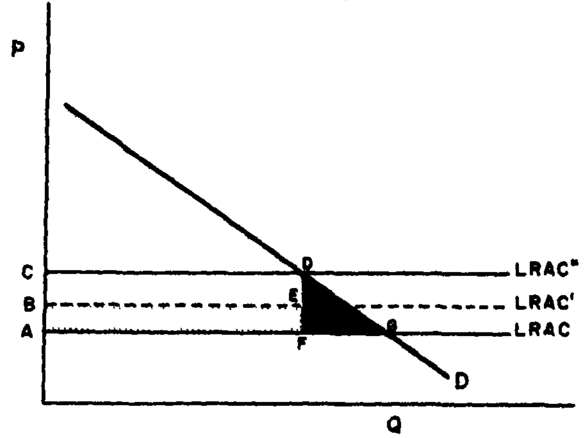


FIG 1 —Welfare effects of sulfur tax

Estimates of welfare gains and losses, which are calculated in the manner suggested by Harberger (1971), are particularly dependent on determination of the appropriate damage function Welfare gains are defined as those resulting from the reduction in sulfur emissions, while welfare losses are the social costs in terms of control costs, losses in consumer surplus, and administrative costs The net welfare gain is the difference between the two components This analysis presumes that sulfur control is the least-cost alternative, for example, neither plants nor residents can relocate at lower costs The administrative costs of the tax are excluded from the calculation both because of inadequate data and because of their insignificant effects on the principal conclusions of this section

Figure 1 depicts the welfare losses resulting from the tax assuming a flat long-run marginal cost function The *LRAC* denotes the long-run average cost function for electricity generation before the imposition of the tax After the imposition of the tax, the new long-run average cost function becomes *LRAC"* as the costs of the taxes, area *BCDE*, and control costs, area *ABEF*, are internalized While the industry's average costs increase by *AC*, the incremental welfare loss consists of the control costs, *ABEF*, and the loss in consumer surplus, *DFG* These sources of welfare loss comprise the two shaded areas in figure 1. Of course, these calculations presume that the full amount of the control costs and taxes are reflected in both residential and industrial electricity prices

Estimates of welfare gains resulting from reduced sulfur emissions are particularly suspect because of the uncertainty surrounding the damage function Following Ridker and Henning's results, we assume that the marginal damage function is flat over the ranges considered in this analysis. Furthermore, a marginal damage cost of $580 per ton of sulfur in 1971 dollars is assumed, as it is the average social cost per ton of sulfur emissions implicit in the Environmental Protection Agency's estimate of $16.1 billion damage from all forms of air pollution.[11] Because of the dubious methodological procedures for this estimate, the reader may want to substitute his own estimate of the marginal damage cost.

Before a comparison of sulfur emission profiles for various tax rates can be made, the cost curves for removing sulfur from fuel oil and coal must be specified. In this context, engineering cost data are a necessary input. The process technology for the desulfurization of fuel oil is well known, and the cost data are reliable. Detailed private estimates[12] of the cost function were used which postulate that the marginal control cost function for fuel oil desulfurization is a step function with two steps.[13]

The technology for economically reducing the sulfur content of coal is not well developed. By the process of wet cleaning, high-sulfur coals can be reduced to 1 percent sulfur, but this appears to have only limited applicability. Flue-gas desulfurization experiments are being conducted with a wet limestone scrubbing process as well as several others, but neither the exact cost nor effectiveness is known. For this reason, the model was simulated under a variety of assumptions, such as "not available" to costs ranging from 14.5¢/10^6 Btu ($3.50 per ton) to 27.1¢/$10^6$ Btu ($6.50 per ton of coal).[14] The most likely case at this point appears to be the simulation at 14.5¢/10^6 Btu ($3.50 per ton). In all simulations, the process is assumed to remove 87 percent of the sulfur. While there is the possibility of opening new low-sulfur coal mines, the average sulfur content of coal before control is assumed to remain constant at its 1964 rate (2.3 percent).[15]

---

[11] See Council on Environmental Quality (1972, pp. 103–7) and Barrett and Waddell (1973).

[12] Private estimates by a large petroleum refining firm are believed superior to public estimates such as those in Hausgaard (1971).

[13] On the first step, the resid desulfurization process can reduce the sulfur content of a barrel of fuel oil from 2.5 to 0.9 percent sulfur at a cost of 42¢/barrel or 8¢/pound of sulfur removed. To reduce the sulfur content of fuel oil from 0.9 to 0.3 percent, vacuum gas oil desulfurization must be utilized at a cost of 14.5¢/pound of sulfur removed. The process technology to reduce sulfur content below 0.3 percent apparently does not exist except by blending low-sulfur crude oils (< 0.3 percent), the supply of which is quite limited. Thus, the marginal cost curve is assumed to become vertical at the point where the sulfur content of fuel oil approaches 0.3 percent.

[14] One of the more recent estimates of the flue-gas desulfurization costs is given in the National Economic Research Associates discussion of electric power in Council on Environmental Quality (1972*a*). According to my rough calculations based on their data, the cost of sulfur removal they are implicitly assuming is $2.30/ton of coal. This is based on taking 7 percent of average revenue per kilowatt-hour (1.7¢/kilowatt-hour), which is 1.2 mills/kilowatt-hour. Of this 1.2 mills/kilowatt-hour, fixed costs for cooling towers and flue-gas processes constitute 0.75 mills/kilowatt-hour, the flue-gas component of fixed costs being 0.6 mills. Combined fixed and variable costs are 0.95 mills/kilowatt-hour (0.6 + 0.35 mills per kilowatt-hour), which implies a fuel cost of 9.5¢/10^6 Btu, or $2.30/ton of coal with standard Btu content. This estimate is believed to be highly optimistic.

[15] Note that no attempt was made to differentiate coals by sulfur content. The U.S. average sulfur content based on U.S. Bureau of Mines (1966, p. 39) was used.

A base case solution to the model was calculated for the period 1972–85 under the assumption that pollution controls or taxes would not be utilized This simulation provides a benchmark by which to evaluate the effects of a sulfur tax The exogenous assumptions and the annual results for key endogenous variables in this simulation are given in Griffin (1974*b*) Of particular interest is the estimated annual emissions of sulfur in millions of tons based on the assumption of no sulfur taxes or controls Despite the rapid growth in nuclear capacity, sulfur emissions increase markedly as generation by fossil-fired plants continues to expand

Table 2 gives the principal results of nine alternative simulations which vary in their assumptions about the tax rate and the cost of flue-gas desulfurization With the exception of the tax adjustment and control cost variants, the same exogenous variables are utilized as in the base case simulation

In all cases, it is assumed that the tax becomes operative in 1976 In addition [illegible] a 15¢/pound sulfur tax, alternative calculations are made for tax rates of 10¢/ and 20¢/pound [16] The cost and sulfur content (after control) for fuel oil desulfurization are similar except in cases 1 and 2, in which, under the 10¢/pound tax, only the use of the resid desulfurization process (which costs 8¢/pound of sulfur removed) is economic

Table 2 reveals that in all cases the cumulative net welfare gains discounted back to 1976 at 7 percent annually are not only positive but of substantial magnitude The magnitude of these numbers is in large part due to the combination of (1) a large estimate of damages ($580 per ton) relative to the tax ($200–$400 per ton), (2) the sustained growth of sulfur oxide emissions in the base case solution, and (3) the effectiveness of the sulfur removal processes The differences between the various cases depend primarily on the availability and cost of flue-gas desulfurization Net welfare gains are the lowest in cases 1, 3, and 6, in which it is assumed that an economically feasible flue-gas desulfurization process is unavailable In the cases in which this process is available, the discounted welfare gains over the 10-year period range from $31 to $42 billion

More important than the absolute size of these net welfare gains is the implication that even if the actual damage rate is one-half the assumed rate of $580 per ton, the net welfare gains in all cases would be positive In fact, if the reader wishes to substitute his own estimate of the incremental damage cost for the EPA estimate, these can be calculated simply

---

[16] The 10¢/pound calculation was made because under the Nixon proposal, some areas that meet the primary standards would qualify for the lower rate, so that on the average the effective tax rate will be between 10¢ and 15¢/pound The 20¢/pound tax calculation is made to conform roughly to the Proxmire proposal, which would begin with a 5¢/pound tax in 1972 and an increment of 5¢/pound each year until it reaches a rate of 20¢/pound in 1975 The Proxmire proposal makes no provision for differential rates depending on the air quality of different geographic areas

TABLE 2

SIMULATION RESULTS FOR ALTERNATIVE TAX RATES AND CONTROL COSTS

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
Assumptions									
Tax rate (cents/pound)	10	10	15	15	15	20	20	20	20
Coal desulfurization (cents/10⁶ Btu)	N.A.	14.5	N.A.	14.5	20.8	N.A.	14.5	20.8	27.1
Sulfur content (%)	2.3	0.3	2.3	0.3	0.3	2.3	0.3	0.3	0.3
Fuel oil desulfurization (cents/10⁶ Btu)	6.6	6.6	11.9	11.9	11.9	11.9	11.9	11.9	11.9
Sulfur content (%)	0.9	0.9	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Results for 1976–85 period									
Discounted cumulative effect (1976–85)									
Welfare gain (10⁹ dollars)	8.59	57.79	12.50	58.84	59.07	14.49	58.88	59.11	59.30
Welfare loss (10⁹ dollars)	0.85	15.77	1.68	16.28	22.16	2.06	16.20	22.05	27.61
Net welfare gain (10⁹ dollars)	7.74	42.02	10.82	42.56	36.91	12.43	42.68	37.06	31.69
Average annual sulfur emission (10⁶ tons)	13.67	2.11	12.70	1.86	1.81	12.23	1.85	1.80	1.75
% sulfur reductions	13.4	86.7	19.6	88.2	88.6	22.5	88.3	88.6	88.9
% due to intrafuel substitution	4.8	79.0	6.1	79.3	76.8	7.0	78.7	76.3	67.2
Average % price change									
Residential	4.1	3.7	6.1	4.0	5.3	8.0	4.1	5.4	6.8
Industrial	5.7	5.1	8.4	5.5	7.3	11.4	5.8	7.6	9.4
% cumulative change in coal use	−10.3	−9.0	−15.2	−9.2	−12.9	−18.6	−9.8	−13.5	−25.4
% cumulative change in fuel oil use	3.7	0.1	12.9	−4.7	5.8	34.5	−4.5	6.0	17.5
% cumulative change in nuclear use	19.5	17.4	29.6	19.1	25.5	29.7	20.5	26.9	30.6

NOTE.—N.A. = not available. All cases assume the same technology but varying tax rates. Regarding flue-gas sulfurization: cases 2, 4, and 7 cost $3.50/ton, cases 5 and 8 cost $5.00/ton, case 9 costs $6.50/ton, flue-gas desulfurization not available for cases 1, 3, and 6.

by adjusting welfare gains proportionately.[17] The new welfare gains can then be subtracted from welfare losses, which are unaffected, to obtain net welfare gains.

In cases involving available flue-gas desulfurization and similar technology, the net welfare gains are only slightly larger for the cases with higher tax rates (e g., cases 2, 4, and 7 and cases 5 and 8) Substitution possibilities other than intrafuel substitution have little effect on sulfur levels in this situation, as their effects are operative on fuels with sulfur contents reduced by 87 percent For example, the substitution of a unit of nuclear generation for a unit of coal generation with only insignificant sulfur content has little effect on sulfur emissions Still another factor contributing to this result is the assumption of homogeneous sulfur content of coal in the model In practice, higher tax rates would force the installation of flue-gas desulfurization for plants burning lower-sulfur coals, so that somewhat higher welfare gains would be expected from the higher tax rates, *ceteris paribus*

In terms of the physical reduction of sulfur emissions, a comparison of the average annual sulfur emissions with the base case simulation indicates marked reductions For example, if an economic flue-gas desulfurization process is available, the sulfur emission level from electric utilities in 1985 may be less than one-fifth of today's level

To demonstrate the point raised earlier with respect to the possibility of substitution between fuels, between capital and fuel, and between other goods and electricity-intensive goods, the percentage of the sulfur reduction attributable solely to the substitution of low-sulfur for high-sulfur fuel of the same type is reported [18] The differences between total sulfur reduction and that due to intrafuel sulfur substitution are attributable to the other forms of substitution frequently overlooked in analyses of this type [19] The importance of these three substitution effects varies significantly depending on the technology In cases 1, 3, and 6, which assume no economic flue-gas desulfurization process, these three substitution effects account for about two-thirds of the change In these cases, fuel oil and nuclear fuel have a decided competitive advantage over coal and induce a substantial interfuel substitution effect This is apparent from the cumulative change in coal, fuel oil, and nuclear use given in the last three rows. In the remaining cases, the three substitution effects are less important, as the reduction in the level of emissions is primarily attributable to intrafuel substitution. This seems plausible, because these substitution effects now have an impact on fuels with a much lower percentage sulfur composition, so that their effects are mitigated

---

[17] For example, if the incremental damage cost is set at $290/ton, the welfare gains in table 2 are reduced by 50 percent. With control costs unchanged, the discounted net welfare gains in billions of dollars equal the following for the nine cases respectively 3 44, 13 13, 4 57, 13 14, 7.38, 5 19, 13.24, 7 51, and 7.60

[18] These were based on the coal and fuel oil purchases in each simulation multiplied by the sulphur differential as a result of control Calculations of this sort are by necessity somewhat arbitrary, as the sum of the partial effects do not necessarily sum to the total, due to interaction effects.

[19] For example, see Hausgaard (1971). For an interesting study incorporating part of these substitution effects, see Chapman (1973)

While the net welfare gain may be the critical variable to the economist, the politician is likely to be more concerned with the distributional effects of a tax Assuming these costs are fully passed forward to residential and industrial users, estimates of the increase in the residential price of electricity range from 3 7 to 8 0 percent For industrial use, price increases vary from 5 1 to 11 4 percent. Despite significant distributional effects, the area representing consumer surplus, *DFG*, is generally small; at most it constitutes $0.4 billion of the $2 06 billion welfare loss in case 6

As would be expected, sulfur taxes have a pronounced effect on fuel oil, coal, and nuclear sales as revealed by comparing cumulative sales with the base case solution. In all cases, nuclear generation will increase, but especially where taxes are high and flue-gas desulfurization is expensive. The petroleum industry stands to gain a marked competitive advantage over coal only in the cases of high tax rates and expensive flue-gas desulfurization (cases 3, 6, and 9) In all cases, coal sales will deteriorate, but the coal industry is well aware, no doubt, of the long-run unfeasibility of the present situation, which is implicit in the base case solution. Thus, from the coal industry's viewpoint, the critical variable is not so much the tax rate but the development of a cheap flue-gas desulfurization process

IV. Appropriate Policy Instruments in a Dynamic Framework

The modeling approach emphasizing dynamic substitution possibilities also gives insights into the selection and effectiveness of air pollution control instruments In the Pigouvian tradition, economists frequently prefer a tax as the optimal control instrument However, sole reliance on a tax requires a detailed knowledge of both the damage and the control cost functions over time. An alternative approach suggested by Baumol and Oates is to utilize both a tax and an ambient air standard This approach may produce suboptimal resource allocation, but it has an administrative cost advantage of requiring no knowledge of the damage or control cost function Baumol and Oates (1971, p. 45) optimistically conclude: "Experience would soon permit the authorities to estimate the tax levels appropriate for the achievement of a target reduction in pollution " Excluding the trivial case of an infinite tax, is it a simple matter to vary taxes in such a manner? Since the control cost function depends on lagged as well as current tax rates, would the agency be required to make substantial rate adjustments over time?

TABLE 3

"PERFECT KNOWLEDGE" SULFUR TAX ADJUSTMENTS
(DOLLARS PER POUND)

Year	ZSG (0%/Year)	1.0%/Year	2.0%/Year
1976	0.15	0.15	0.15
1977	16.70	11.11	6.03
1978	11.08	8.04	5.72
1979	3.86	0.15	0.15
1980	8.01	3.00	0.15
1981	11.27	6.74	0.15
1982	12.47	6.70	0.15
1983	14.34	6.77	0.15
1984	11.73	0.15	0.15
1985	16.69	2.59	0.15
1986	21.96	15.80	0.15
1987	13.86	0.15	0.15
1988	13.76	0.15	0.15
1989	21.03	2.61	0.15
1990	23.22	8.32	0.15

Suppose, after initiating a 15¢/pound sulfur tax, that in 1976 society selects a standard of a zero sulfur emission growth, *ZSG*, or some variant of it, and chooses to use annual tax variations to achieve this air standard. The model is then solved for the necessary tax adjustments assuming the regulatory agency has "perfect knowledge" of the control cost function. This simulation assumes exogenous assumptions similar to those used in table 2 along with the initial condition of a 15¢/pound tax in 1976, and the technology implicit in case 4 of table 2. The tax rate is endogenized.

Table 3 illustrates the necessary variations in taxes for *ZSG*, for a 1 percent/year emission growth, and for a 2 percent/year sulfur emission growth.[20] The results indicate that substantial year-to-year modifications would have to be made in the tax rate and that the more ambitious the standard, the greater the tax adjustments necessary. The relatively large magnitude of the tax adjustments required should not be taken too seriously, because the standards selected are quite ambitious by any measure.[21] Furthermore, at high tax rates, one would expect greater substitution possibilities than those implicit in a model estimated over a much smaller range of price variations. New process technology in the form of sulfur removal equipment (which is exogenous to the model) also could be expected to develop in response to the above tax rates, thereby reducing the tax adjustments necessary.

---

[20] The sample period simulations were extended to 1990 to illustrate that the cyclical phenomena appear to continue even over the longer period. The exogenous values were simply extrapolated for the period 1985–90.

[21] The initial target in 1976 is set at 1.63 million tons of sulfur emissions compared with an estimated 13.1 million tons of sulfur in the base case solution without air pollution controls in 1976.

The most interesting result is that the necessary tax adjustments exhibit a *cyclical behavior* which is not apparent a priori The cyclical variations arise from the fact that lagged electricity prices and lagged relative fuel prices in the model have a greater impact on current sulfur emissions than the current period's prices, which include the effect of the current tax rate. Consequently, the problem with using tax rates as a short-run policy instrument is much more complex than the case in which there are no long-run effects, and tax rates can be adjusted continuously until equilibrium is reached The problem becomes one of dynamic optimization rather than static optimization

Several more general implications emerge from the preceding exercise First, the traditional static approach to pollution questions by economic theorists is likely to ignore the inherently dynamic nature of the problem This result also has important policy implications The sale of transferable pollution licenses[22] is likely to be a more efficient (less expensive in terms of information requirements) approach to obtaining target quantities of pollution control than the Baumol and Oates approach Presumably, market forces would produce prices more similar to those in table 3 than those set by an agency with imperfect knowledge as to the lagged substitution effects Another interpretation of these results is to raise the question of whether the benefits of obtaining short-run pollution targets exceed the costs Even if the Dales proposal generated ticket prices approximating the tax rates in table 3, such wide cyclical variations in the tax rates may periodically generate substantial excess capacity of pollution control equipment, thereby creating offsetting welfare losses The problems with using taxes to achieve year-to-year changes in sulfur emissions are in many respects analogous to those faced in stabilization policy However, in the case of stabilization policy, there are significant social benefits to correct policy, while the welfare benefits of meeting some arbitrary pollution standard are unclear, if possible While the use of taxes to achieve year-to-year adjustments in pollution levels seems dubious, it does not follow that tax policies designed to hold taxes at a constant rate for long periods are also suspect To the contrary, the results of table 2 indicate that appropriate tax rates held at sustained levels serve as a very efficient policy instrument

V. Other Factors

The preceding aggregate analysis provides useful general guidelines, nevertheless, there are several issues which should be considered in formulating an optimal policy By aggregating over space and over the period of 1 year, the model may omit potentially important geographical and seasonal effects. In practice, a tax rate should depend primarily on regional factors rather than on some national rate applicable to all regions. For example, the relationship between quantity of sulfur emissions and air quality measures may vary substantially across regions for climatological reasons. The marginal damage function depends on population densities, which vary with geographic region. Regional differences in per capita income are likely to affect the damage function, as pollution abatement is a superior good Also, tastes probably differ between regions, depending in large part on individuals' economic interdependence with the local polluting industry. Thus, there are a number of variables, such as climatological factors, population densities, income levels, and tastes, which are likely to produce geographically distinct damage functions. Despite these geographical differences, present proposals app[illegible] to give at most only modest recognition of these problems They call for either a uniform tax rate (e.g , Proxmire proposal) or a differential tax rate (e g , Nixon proposal), which would allow for a 10¢/pound tax in regions where the primary air quality standard is met and a zero tax in regions where the secondary air quality standard is met In addition, if the marginal damage function is not horizontal, some consideration should be given to differential tax rates, depending on the seasonal variation in electricity output With the increasing peak load electricity requirements in the summer, tax rates may have to be raised to reflect higher summer damage costs

---

[22] For a further discussion by an early advocate of this approach, see Dales (1968, pp. 92–97).

Following Harberger (1971), one might attempt to approximate second-best effects by adjusting the net welfare gains in table 2 to include the welfare gains or losses associated with distortions in industries related to electricity Wilson (1971) and Anderson (1972) show a small but significant cross-elasticity of demand between natural gas and electricity pointing to one source of distortion The administered pricing system for natural gas appears to have resulted in a situation in which at current consumption levels the long-run supply price exceeds the administered price A sulfur tax will increase the price of electricity, thereby shifting outward the demand schedule for natural gas, raising production in the short run, and exacerbating the welfare loss Likewise, if the environmental costs of nuclear generation are not internalized, sulfur taxes will increase nuclear generation and produce some additional welfare loss This paper will not attempt to estimate these additional losses, because they are not likely to change the principal conclusion of table 2—that the imposition of sulfur taxes should result in a net welfare gain

Nevertheless, the additional welfare loss does have implications for the tax rate chosen A comparison of net welfare gains between tax rates of 15¢ and 20¢/pound for similar technology (case 3 with 6, case 4 and 7, case 5 with 8) indicates only a slight preference for the 20¢/pound tax.

If the likelihood of the technical assumptions in case 9 is low, then a strong argument could be made in favor of the 15¢/pound rate due to the second-best effects. Similarly, if flue-gas desulfurization were certain to be available at $3.50 per ton of coal or less, then a 10¢/pound tax (case 2) would be preferable for similar reasons.

VI. Summary

The extension of econometric modeling to problems of environmental policy offers a potentially fruitful and relatively unexplored area of research The dynamic framework, which includes various substitution possibilities, appears to offer some useful policy insights despite a number of restrictive assumptions Over a wide range of marginal damage rates and technological assumptions, the Pure Air Tax Act of 1972 should result in a net welfare gain However, it does not follow that the proposed tax rate is optimal, as the preceding section raises a number of unanswered questions The pervasive nature of the dynamic substitution effects implicit in the model suggests skepticism with proposals using tax adjustments to maintain some given air quality standard on a year-to-year basis Much more serious thought must be given the issue of selecting the appropriate policy instruments in a dynamic context

References

Anderson, Kent P "Residential Demand for Electricity Econometric Estimates for California and the United States" RAND Report R-905-NSF Santa Monica, Calif RAND Corp , January 1972

Anderson, R. J , Jr , and Crocker, T D "Air Pollution and Housing Some Findings" Paper no 264 Lafayette, Ind Inst Res Behavioral Econ. and Management Sci , Purdue Univ , January 1970

Ayres, R U , and Kneese, A V "Production, Consumption, and Externalities." *A E R* 59 (June 1969) 283–97

Barrett, L B , and Waddell, T E "Cost of Air Pollution Damage A Status Report" Washington Environmental Protection Agency, February 1973.

Baumol, W. J , and Oates, W. E "The Use of Standards and Prices for Protection of the Environment." *Swedish J Econ.* 73 (March 1971) 42–54

Chapman, D. "Internalizing and Externality A Sulphur Emission Tax and the Electric Utility Industry" Paper presented at the February 1973 M.I T Conference on Energy

Congressional Record. Vol 118 Proxmire Bill (S 3057), January 24, 1972, S 276; Nixon Proposal, February 8, 1972, H 841, 92d Cong , 2d sess., 1972

Council on Environmental Quality *The Economic Impact of Pollution Control* Washington Environmental Protection Agency, March 1972. (*a*)

———. *Environmental Quality, 1971* Washington· Environmental Protection Agency, March 1972 (*b*)

Crocker, T. D. "Urban Air Pollution Damage Functions Theory and Measurements." Final Report to the National Air Pollution Control Administration, CPA 22-69-52 Madison, Wis , 1971

Dales, J H *Pollution, Property, and Prices.* Toronto: Univ. Toronto, 1969.

Griffin, J. M "The Effects of Higher Prices on Electricity Consumption " *Bell J.* (1974), in press (*a*)

——— "Energy Use and Environmental Quality A Look at Prospects and Policy Alternatives " *In Studies in Taxes and Energy,* edited by Gerard Brannon Boston· Ballinger, 1974, in press (*b*)

Harberger, Arnold C. "The Basic Postulates for Applied Welfare Economics An Interpretive Essay." *J Econ Literature* 9 (September 1971) 785–97

Hausgaard, O "Proposed Tax on Sulphur Content of Fossil Fuels" *Public Utilities Fortnightly,* September 16, 1971, pp 27–32

MacAvoy, Paul W *Economic Strategy for Developing Nuclear Breeder Reactors* Cambridge, Mass M I T., 1969

Preston, Ross *The Wharton Annual and Industry Forecasting Model* Philadelphia Univ Pennsylvania, 1972

Ridker, R G *Economic Cost of Air Pollution* New York Praeger, 1967

Ridker, R. G , and Henning, J A "The Determinants of Residential Property Values with Specific Reference to Air Pollution " *Rev Econ and Statis* 49 (May 1967): 246–57

U S., Bureau of Mines *Sulphur Content of United States Coals* Information Circular 8312 Washington Bur Mines, 1966.

U S , Congress, Senate, Subcommittee of the Committee on Public Works *Hearings on Air Pollution,* 1970, pt 1, 91st Cong , 2d sess , 1970

Wieand, K F "Property Values and the Demand for Clear Air Cross-Section Study for St Louis " Ph D dissertation, Washington Univ , 1970

Wilson, John W "Residential Demand for Electricity " *Q Rev Econ and Bus* 2 (Spring 1971) 7–22

Searching for the Lowest Price When the Distribution of Prices Is Unknown

---

Michael Rothschild

Princeton University

Search theory, which purports to explain how individuals behave when they have imperfect or incomplete market information, has received much attention recently Economists have derived a number of results characterizing the effects of various changes on optimal-search behavior Almost without exception, these results depend on the untenable assumption that searchers know the probability distribution from which they are searching This paper studies the effect of assuming instead that searchers learn about the probability distribution while they search from it Not invariably, but in many instances, the qualitative properties of optimal-search strategies—and thus the behavior of those who follow them—are the same as in the simpler case when the distribution is assumed known

I. Introduction

Stories about search occupy a central position in the "new microeconomics of inflation and employment " This accounts for much of the recent work by economists[1] analyzing variants of the following problem A man is considering purchase of some good which is sold at different

A preliminary version of this paper was presented at the NBER conference on stochastic control at the University of Chicago, June 1973 I benefited from discussions with David Blackwell, Gregory Chow, Stanley Fischer, Emmett Keeler, Robert E Lucas, Daniel McFadden, Roy Radner, and Yoram Weiss This research was supported by the Industrial Relations Section of Princeton University, the Mathematical Social Science Board, and the National Science Foundation

[1] See Stigler (1961, 1962), McCall (1965, 1970), Nelson (1970), Gastwirth (1971), Gronau (1971), Gould (1972), Salop (1973), Telser (1973), Whipple (1973), and Kohn and Shavell (1974)

stores at different prices. He can elicit price quotations from the various sellers by paying a fee.[2] What search strategy will the man follow?

Economists are interested in the rules which searchers follow, because these rules determine the demand functions that sellers in such markets face and thus, in part, the nature of the markets themselves. Although the only way to settle the question of what rules searchers follow is by observation, very little empirical work has been done on this problem.[3] Instead, researchers have proposed search rules that reasonable consumers might follow and have examined their properties. Stigler, who is largely responsible for introducing this topic into economic theory, suggested that the individual should visit n stores, obtain price quotations from each one, and then buy from the lowest-price store. The expected price paid will be

$$M_n = \int_0^{\infty} [1 - F(p)]^n \, dp,$$

where $F(p)$ is the distribution of prices. If such a procedure is followed, the only decision variable is n, the number of stores visited. Stigler pointed out that the familiar marginal calculations of microeconomic theory will suffice to determine n. Clearly, M_n is a decreasing function of n, while the expected gain from searching,

$$G_n = M_{n-1} - M_n = \int_0^{\infty} [1 - F(p)]^{n-1} F(p) \, dp,$$

also decreases with n. Therefore, n should be chosen so that

$$G_n \geq c > G_{n+1}.$$

This rule has several interesting properties.

1. If all its potential customers follow this rule, then a firm faces a well-behaved demand function; expected sales are a nonincreasing function of the price it charges.

2. Customers' search behavior is a function of the cost of search c and the distribution of prices $F(\ )$. Thus, it is possible to do comparative

---

[2] This fee, the cost of search, is generally interpreted as the cost of visiting a store and obtaining a price quotation. It could also include the cost of doing without the item in question while search continues, in which case the fact that search takes time would have to be formally incorporated in the model (see n. 10 below). Other rather trivial, from the formal point of view, generalizations are possible. Price and cost may be measured in utility rather than money. However, the utility function must be linear, so this is a small generalization. The same framework may be used to analyze search for jobs when wages (or the utility of jobs) are random and unknown. I shall largely ignore these matters of interpretation.

[3] A cynic might suspect that the lack of empirical work reflects a lack of interest in the real world on the part of those concerned with search rules. A glance at the authors in n. 1 above suggests that this is not the only reason. It seems to me—and my interest in such matters is admittedly casual—that the problem of determining characteristics of search rules empirically is a very difficult one.

statics so as to examine the consequences changes in $F(\ )$ and c The most important results of these exercises are:

3. If costs of search increase, the amount of search decreases

4. As prices become more dispersed, expected total costs decrease. This follows from the fact that, for all n, M_n decreases as the distribution becomes more variable.[4] Other things equal, customers prefer to draw from riskier distributions This preference for risk or uncertainty on the part of those normally deemed risk averse can explain apparently odd or perverse phenomena. For an application to the theory of migration, see David (1973)

Interesting as these results are, they depend on people following a particular search rule—and not a particularly attractive one at that Although fixed-sample-size rules have a certain intuitive appeal, they are not the best search procedures and are in some circumstances simply silly A person who rigidly follows a fixed-sample-size rule will, even if he gets a price quotation less than the cost of search, keep on sampling until his quota of price quotations is fulfilled It is thus comforting to know that the optimal-search rule has all the attractive properties of the fixed-sample-size rule The optimal rule is sequential (after receiving each price quotation, the searcher decides whether to continue searching or to accept the quoted price) and is characterized by a reservation price; there is a price R such that the searcher will accept any price less than or equal to R while he will reject a price higher than R Once again, familiar economic reasoning suffices to determine R If the lowest price the cus-

---

[4] This is generally demonstrated by calculating $(dM_n/d\sigma^2) < 0$ for particular parametric families of distributions However, the result is much more general Let $F(p, t)$ be a family of distribution functions with support in $[0, B]$ indexed by t Then, as Diamond and Stiglitz (1974) have shown, increases in t correspond to increases in risk in the sense of Rothschild and Stiglitz (1970) if

$$\int_0^B dF(p, t) = 1, \tag{i}$$

$$\int_0^Y F_t(p, t)dp \geq 0, \qquad 0 \leq Y \leq B, \tag{ii}$$

and

$$\int_0^B F_t(p, t)dp = 0 \tag{iii}$$

If $h(p)$ is any increasing function, then

$$\int_0^B h(p)F_t(p, t)dp \leq 0$$

Note that

$$M_n(t) = \int_0^B [1 - F(p, t)]^n dp$$

and

$$\frac{dM_n(t)}{dt} = \int_0^B -n[1 - F(p, t)]^{n-1} \ F_t(p, t)dp \leq 0,$$

since $-n[1 - F(p, t)]$ is increasing in p.

tomer has received to date is S, then the expected gain from searching once more is

$$g(S) = \int_0^S (S - p)\, dF(p) = \int_0^S F(p)\, dp.$$

The optimal rule is to search whenever the cost of an additional search is less than the expected gain from that search so that R must satisfy

$$g(R) = c \tag{1}$$

This reservation-price rule has the four properties listed above That properties 1 and 2 hold is obvious Increasing c increases the reservation price and thus lowers the intensity of search so that 3 holds. Kohn and Shavell (1974) have shown that increasing dispersion lowers the reservation price and thus total expected costs[5] so that 4 holds for this rule as well They do this by showing that if $F(p, t)$ is, as in note 4 above, a family of price distributions which becomes more dispersed as t increases, then

$$g(S, t) = \int_0^S F(p, t)\, dp$$

is an increasing function t Therefore, if

$$c = g(R, t) = g(R', t')$$

with $t' > t$, then $R' < R$ Thus, the reservation price decreases as price variability increases, or

5 Increased price dispersion increases the intensity of search This last result—which does not hold in general for fixed-sample-size rules[6]—is a kind of stability property which could be used in a complete model (one which explained price distributions endogenously) to show the existence of an equilibrium distribution of prices It should turn out in most sensible models that increased search activity will decrease price dispersion [7]

These results depend on the assumption that the searcher behaves as if he knows the distribution of prices In any economic context, this is a very bad assumption Little is known about the nature of price distributions, and it seems absurd to suppose that consumers know them with any degree of accuracy.[8] Since the major reason for believing that searchers follow optimal sequential rules is that they *are* optimal, it is important that their cost-minimizing properties not depend crucially on their being based on correct knowledge of the price distribution This is unfortunately not true Gastwirth (1971) explored the robustness of optimal-reservation-price rules. He found that modest specification errors could lead to dramatic increases in the expected number of searches and in the expected cost of buying For example, someone who chose a reservation price on the assumption that prices were distributed uniformly on the unit interval $(F(x) = x)$ when they really were distributed according to a right-triangular distribution on the same interval $(F(x) = x^2)$ would on the average incur roughly twice the total costs and search five times as much as he would if he were correctly informed If results 1–5 are to be salvaged, the problem of what the searcher should do if he does not know the price distribution must be attacked This can be done in two ways by exploring the properties of reasonable rules of thumb, or by characterizing optimal rules Telser (1973) took the first approach He calculated (using Monte Carlo techniques) the expected costs of various search rules—which had properties 1–4—against several differently shaped price distributions and compared these costs with those which would ensue if the searcher followed the naive rule of taking the first price offered him

---

[5] Total expected costs to a searcher following a reservation-price rule are just equal to the reservation price This follows from the fact that expected costs do not change as the searcher continues to sample The logic of rule (1) might be stated as Accept any price less than the expected costs of continuing to search, otherwise continue

[6] If, in the notation of n 4, $G_n(t) = M_{n-1}(t) - M_n(t)$, then it is not true that $G_n(t)$ increases with t whenever $F(p, t)$ satisfies (i), (ii), and (iii) I suspect that most increases in price dispersion which do not lead to piling up of probability mass at the extreme points of the distribution will increase $G_n(t)$, but I have not yet been able to formulate this precisely

[7] For a discussion of the need for such a complete model, see Rothschild (1973)

This paper takes the second tack Optimal-search rules from unknown distributions are derived and characterized [9] The results of this exercise are as follows In Section II the problem of a man who knows prices belong to some finite set, but does not know how they are distributed, is formally described Section III discusses an important example—the case where the prior distribution is a Dirichlet In the next section, dynamic programming is used to derive the optimal strategy It is possible to parameterize the problem so that the optimal-valuation functions are continuous—a fact which is exploited in Section V, where it is shown in Theorem I that if a person follows the optimal strategy, search terminates after a finite number of searches This is used to prove that property 3 holds for search from unknown distributions, that is, that search decreases as its cost increases. In Section VI it is shown (in Theorem II) that if prior beliefs are Dirichlet the optimal-search rule has a reservation-price property. Searchers will accept a price if and only if it is less than some particular price p_R The reservation price is a function of the searcher's beliefs, it changes as his beliefs are revised in the light of experience. Section VII is devoted to proving, in Theorem III, an analogue of properties 4 and 5 for the Dirichlet case

---

[8] I suspect they vary a great deal Characteristics of empirical wage distributions—which are related to (but distinct from) distributions of wage offers—are unstable over occupations and across cities (Buckley 1969)

[9] The statistical literature on optimal-search rules is extensive Breiman (1964), DeGroot (1970, chap 13), and Chow, Robbins, and Siegmund (1971) are surveys of increasing sophistication Some of the results of this paper appear to be new, particularly Theorems I, II, and III The work most closely related to that presented here is that of Yahav (1966) and DeGroot (1968) Yahav considers the general problem of searching with recall (the searcher is allowed to return to stores previously sampled) from an unknown distribution, shows that optimal rules exist under quite general circumstances, and discusses how they could be approximated DeGroot completely solves the problem faced by a man searching (with or without recall) from a normal distribution of known variance and unknown precision

The most important results of this paper are that for the example of Section III, optimal-search rules from unknown distributions have the same qualitative properties as optimal rules from known distributions Since it is easy to construct examples for which this is not true, it is natural to ask how general these results are The final section considers this question. I believe them to be more general than it may appear, as they do not seem to depend on prices being confined to a finite set However, they are still quite special, as the proofs depend on the process of revising beliefs to accommodate new information having a particular—and not terribly natural—local property Whether similar results hold if this assumption is abandoned is an open and difficult question

II. Preliminaries

Consider a man trying to buy a good at the lowest total expected cost At the beginning of each period, this man pays c and receives a price quotation, which upon receiving he must decide whether to accept or to pay (at the beginning of the next period) c again and receive another price quotation Total costs include the price actually paid as well as the costs of search [10] I assume that the man does not have the privilege of recall, that is, that an offer once spurned cannot be taken up again As is well known, this assumption is innocent—it has no effect on the optimal strategy—when the price distribution is assumed known It is not innocent in the present case

For simplicity, suppose that there are only a finite number of prices $p_1, p_2, \ldots, p_n$ and label them in ascending order so that

$$p_i < p_{i+1} \tag{2}$$

The probability distribution of prices is a multinomial distribution It is completely characterized by the vector Π whose ith element π_i is the probability that the ith price is chosen Since Π is a probability distribution,

$$\Pi \in \Delta = \{(x_1, \ldots, x_n) \in R^n \mid x_i \geq 0, \sum x_i = 1\}$$

---

[10] Given that the process takes place in time, perhaps it would be more logical to minimize discounted expected costs This would introduce no serious complications See Kohn and Shavell (1974) for a discussion of the effect of varying the discount rate on search rules.

Previous work has assumed that the searcher knows Π. Instead, I assume that he has a prior distribution $F(\ )$ over Δ. As he continues to search, he gathers more information about the distribution of prices which he assimilates by updating his prior according to Bayes's rule. All necessary information about his experience is contained in the statistics

$$N = (N_1, \ldots, N_n),$$

where N_i is the number of times price i has been observed.

It is convenient to parameterize his experience slightly differently. If $S(N) = \sum_i N_i$, then the vector whose ith element is

$$\mu_i = \frac{N_i}{S(N)} \tag{3}$$

represents the average number of times that each price has been observed, while

$$\rho = S(N)^{-1} \tag{4}$$

represents the total number of prices he has observed, therefore, (μ, ρ) also contains all the information the searcher has accumulated. This parameterization permits a distinction between the content of this information, represented by μ, and its precision, represented by ρ. With these definitions, new information is assimilated as follows. With the observation of price i, (μ, ρ) becomes

$$h_i(\mu, \rho) = \left(\frac{\mu_1}{\rho + 1}, \ldots, \frac{\mu_i + \rho}{\rho + 1}, \ldots, \frac{\mu_n}{\rho + 1}, \frac{\rho}{\rho + 1}\right) \tag{5}$$

There is a slight technical difficulty with this convention. The updating rule (5) is not consistent with the definitions (3) and (4) when $N = 0$. This problem may be finessed. Suppose $X = (x_1, \ldots, x_n)$ and $t = (t_1, \ldots, t_n)$, define

$$X^t = x_1^{t_1} x_2^{t_2} \cdots x_n^{t_n} \tag{6}$$

A man with initial beliefs $F(\Pi)$ and no experience is no different from a man with initial beliefs $\hat{F}(\Pi)$ and experience $N = (N_1, \ldots, N_n)$, where $\hat{F}(\Pi)$ is a probability distribution over Δ satisfying $\hat{F}(\Pi)\Pi^N = KF(\Pi)$ for some positive constant K or, using the definitions (3) and (4),

$$\hat{F}(\Pi)\Pi^{(\mu/\rho)} = KF(\Pi)$$

In the sequel, we largely ignore the prior distribution $F(\ )$ and focus instead on the information

$$(\mu, \rho) \in \Gamma = \Delta \times [0, 1], \tag{7}$$

which is updated according to (5).

Corresponding to any (μ, ρ) is a vector $\lambda(\mu, \rho)$ whose ith component is given by

$$\lambda_i(\mu, \rho) = \frac{\int_\Delta \pi_i \Pi^{\mu/\rho}\, dF(\Pi)}{\int_\Delta \Pi^{\mu/\rho}\, dF(\Pi)} \tag{8}$$

The term $\lambda(\mu, \rho)$ is a probability distribution which represents the searcher's expected beliefs in that he would take a small bet on the proposition that the next price observed would be p_i at the odds $\lambda_i(\mu, \rho)$ to $[1 - \lambda_i(\mu, \rho)]$ Since $\lambda(\mu, \rho)$ may be considered the index of a posterior distribution based on a sample size of ρ^{-1} which converges to a normal distribution with mean equal to the sample mean as sample size increases,

$$\lim_{\rho \to 0} \lambda_i(\mu, \rho) = \mu_i \tag{9}$$

III. An Example

A simple and important example will illustrate the nature of the searcher's problem and clarify the meaning of N, (μ, ρ), and $\lambda(\mu, \rho)$ Suppose that the searcher's prior is a Dirichlet distribution Since the Dirichlet is the conjugate prior of the multinomial distribution, the posterior distribution will also be a Dirichlet The Dirichlet is an n parameter distribution, completely characterized by the numbers $N = (N_1, \ldots, N_n)$[11] The properties of the Dirichlet are best illustrated by the following parable, which describes a problem exactly equivalent to that faced by the searcher[12] Assume the N_i are integers In an urn there are $S(N) = \sum N_i$ pieces of paper, p_1 is written on N_1 of these slips, p_2 on N_2 of them, and so on A man draws from the urn at random If he draws a slip with p_i on it he may either pay p_i or pay c, return the slip to the urn, place another slip with p_i on it in the urn, and draw from the urn again That is, if he chooses to continue sampling, he faces the same problem as before except that the parameter describing the urn is

$$J_i(N) = N + e_i, \tag{10}$$

where e_i is the vector with 1 in the ith place and 0 elsewhere The probability of getting price p_i from an urn with parameter N is just

$$\lambda_i(N) = \frac{N_i}{S(N)}$$

---

[11] See, for example, DeGroot (1970, pp 49–51) for a discussion of the properties of the Dirichlet distribution

[12] I am indebted to David Blackwell for pointing out to me this way of looking at Dirichlet distributions

The quantity $S(N)$ measures how fast these probabilities will change with successive drawings. If $S(N)$ is small, new drawings will alter the composition of the urn considerably; if $S(N)$ is large, they will hardly affect it. There is no reason why the N_i have to be integers. The above problem is equivalent to the searcher's problem for any positive N as long as the probability of drawing p_i from an urn with parameter N is $N_i/S(N)$ and N is updated by (10) when p_i is observed. Letting $\mu_i = N_i/S(N)$ and $\rho = S(N)^{-1}$, it is easy to check that the updating formula (10) is consistent with (5) and that $\lambda_i(\mu, \rho) = \lambda_i(N) = N_i/S(N) = \mu_i$.

It is important to realize that this story does not describe how price quotations are generated. In fact, there is a real price distribution which generates price quotations. However, the searcher does not know this distribution. His knowledge of the distribution is described by N, the composition of the urn. If price quotations are really generated by Π, then the strong law of large numbers states that the proportion of prices in the urn will, with probability 1, eventually be equal to Π.

The Dirichlet case is more general than it might appear. I have shown that if the prior is Dirichlet then it is possible to parameterize experience so that $\lambda_i(\mu, \rho) = \mu_i$. As I show below, this property characterizes the Dirichlet completely. Thus, it follows from (9) that as experience accumulates, all searchers come to behave as if their priors were Dirichlet.

Proposition 1. A searcher has a Dirichlet prior if and only if it is possible to parameterize his experience so that $\lambda_i(\mu, \rho) = \mu_i$ for all $(\mu, \rho) \in \Gamma$.

Proof: The "only if" part of the proposition has already been demonstrated. Suppose $F(\ )$ is a probability measure on Δ and that, when $\lambda_i(\mu, \rho)$ is defined by (8), $\lambda_i(\mu, \rho) = \mu_i$. There is a Dirichlet distribution $\tilde{F}(\ )$ so that, when $\tilde{\lambda}(\mu, \rho)$ is defined analogously, $\tilde{\lambda}(\mu, \rho) = \lambda(\mu, \rho)$. It follows that for all $N \geq 0$

$$\frac{\int_\Delta \Pi^{N+e_i}\, dF(\Pi)}{\int_\Delta \Pi^{N}\, dF(\Pi)} = \frac{\int_\Delta \Pi^{N+e_i}\, d\tilde{F}(\Pi)}{\int_\Delta \Pi^{N}\, d\tilde{F}(\Pi)} \tag{11}$$

It is easy to see by induction that

$$\int_\Delta \Pi^N\, dF(\Pi) = \int_\Delta \Pi^N\, d\tilde{F}(\Pi)$$

for all nonnegative integer N. The Stone-Weierstrass theorem implies that

$$\int_\Delta u(\Pi)\, dF(\Pi) = \int_\Delta u(\Pi)\, d\tilde{F}(\Pi)$$

for all bounded continuous functions, $u(\ )$, defined on Δ. Thus, by theorem 1.3 of Billingsley (1968), $F(\ )$ and $\tilde{F}(\ )$ coincide.

IV. The Optimal Strategy

It is now possible to describe the optimal strategy of a man whose knowledge of prices is represented by the parameters (μ, ρ) which are updated according to (5) This is done in the standard way, by induction. Let

$$V_0(\mu, \rho) = \sum_i \lambda_i(\mu, \rho) p_i \tag{12a}$$

and

$$V_T(\mu, \rho) = \sum_i \lambda_i(\mu, \rho) \min \{p_i, V_{T-1}[h_i(\mu, \rho)] + c\} \tag{12b}$$

The quantity $V_{T-1}(\mu, \rho)$ is the minimum expected cost incurred by a searcher with prior experience (μ, ρ) who is allowed to search at most T times but must accept the Tth price offer made to him It is easy to see by induction that $V_T(\mu, \rho) \leq V_{T-1}(\mu, \rho)$ and $V_T(\mu, \rho) \geq p_1$ for all T, so that the $V_T(\mu, \rho)$ converge Let

$$V(\mu, \rho) = \lim_{T \to \infty} V_T(\mu, \rho)$$

Then $V(\mu, \rho)$ satisfies

$$V(\mu, \rho) = \sum_i \lambda_i(\mu, \rho) \min \{p_i, V[h_i(\mu, \rho)] + c\}, \tag{13}$$

an equation which defines the optimal policy If p_i is drawn when beliefs are (μ, ρ), accept if

$$p_i \leq V[h_i(\mu, \rho)] + c, \tag{14}$$

otherwise elicit another price offer

Proposition 2 $V(\mu, \rho)$ is continuous

This fact, although of little interest itself, is the basis of the proof of the next section and justifies the use of extreme examples in the sections that follow It is important to realize that all the proofs given hold for any $(\mu, \rho) \in \Gamma$ including such strange boundary values as $(\mu = e_i, \rho = 0)$ or $(\mu = e_i, \rho = 1)$, where e_i is the vector whose ith component is 1 while all other components are equal to 0

Proof: Since $\lambda_i(\mu, \rho)$ is continuous, so is $V_0(\mu, \rho)$ A simple induction establishes that $V_T(\mu, \rho)$ is continuous for all T It remains to show that the V_T converge uniformly

Let $V_T^+(\mu, \rho)$ be the expected total cost to a searcher who for T periods follows the policy of accepting the elicited price only if (14) is satisfied, if after T periods no price has been accepted, the searcher must pay p_n The quantity $V_T^-(\mu, \rho)$ is defined similarly except that if he has not chosen after T periods he receives p_1 From their definitions it follows that

$$V_T^+(\mu, \rho) \geq V_T(\mu, \rho) \geq V(\mu, \rho) \geq V_T^-(\mu, \rho).$$

It will suffice to show that $|V_T^+(\mu, \rho) - V_T^-(\mu, \rho)|$ converges to 0 uniformly:

$$|V_T^+(\mu, \rho) - V_T^-(\mu, \rho)| \leq \beta_T(\mu, \rho)p_n,$$

where $\beta_T(\mu, \rho)$ is the probability—calculated according to the prior (μ, ρ)—that sampling will not terminate after T periods. However, from the definition of $V_T(\mu, \rho)$ it must be that

$$\beta_T(\mu, \rho)Tc \leq V_T(\mu, \rho) \leq p_n,$$

so $\beta_T(\mu, \rho) \leq p_n/Tc$ and $\beta_T(\mu, \rho)$ converges to 0 uniformly. This completes the proof.

V. Finiteness of Search

In this section it is shown that search will cease after a finite time. The idea of the proof (for which I am indebted to Robert Lucas) is a straightforward one. Regardless of his beliefs, a person will always accept p_1. If p_1 does not appear for a long time, then he will come to believe that p_2 is the lowest possible price and will of course accept it. Thus, there is a t_2 such that after t_2 trials either the person has stopped sampling or he will be willing to accept p_2. Similar arguments establish the existence of $t_3, t_4, \ldots, t_n$ with the same property. After t_n trials, the searcher has either stopped or will accept any price that occurs.

Theorem I: There is a number t such that a person following the optimal strategy will have stopped searching after t trials.

Proof. The theorem is *not* a probability statement. Let Ω be the set of all infinite sequences of the prices $p_1, \ldots, p_n$ and let ω denote an element of Ω. I shall show that there is a t such that for any ω sampling will have stopped by the time the tth price is quoted. Before I do so, some notation is necessary. For any ω, let ω_s denote its sth coordinate. Since (μ, ρ) are updated as functions of ω, we may define $(\mu, \rho)[\omega(q)]$ as the values of (μ, ρ) obtained after observation of $\omega_1, \ldots, \omega_q$.

Define the acceptance set A_i by $A_i = \{(\mu, \rho) \in \Gamma \mid p_i < V[h_i(\mu, \rho)] + c\}$. It follows from (9) that if $K_2 = \{(\mu, \rho) \in \Gamma \mid \mu_1 = \rho = 0\}$, then $K_2 \subset A_2$. Since A_2 is open and Γ compact, there is an $\varepsilon > 0$ such that if $(\mu_1, \ldots, \mu_n, \rho) \in \Gamma$ and $\mu_1 + \rho < \varepsilon$, then $(\mu_1, \ldots, \mu_n, \rho) \in A_2$. We may choose t_2 such that if

$$\omega_s \neq p_1 \qquad \text{for } s < t_2, \tag{15}$$

then $\mu_1[\omega(t_2)] + \rho[\omega(t_2)] < \varepsilon$, thus if ω satisfies (15), $(\mu, \rho)[\omega(t_2)] \in A_2$. Furthermore, if ω satisfies (15) and $\omega_{s_1} = p_2$ for $s_1 > t_2$, then sampling will have terminated by s_1, since either $\omega_s = p_1$ for $t_2 < s < s_1$ or $(\mu, \rho)[\omega(s_1)] \in A_2$.

Now let $K_3 = \{(\mu, \rho) \in \Gamma \mid \mu_1 = \mu_2 = \rho = 0\}$. Clearly $K_3 \subset A_3$ and, by arguments analogous to those used above, there exists a t_3 such that if ω satisfies (15) and

$$\omega_s \neq p_1, \qquad w_s \neq p_2 \qquad \text{for } t_2 \leq s < t_3,$$

then

$$(\mu, \rho)[\omega(t_3)] \in A_2 \cup A_3$$

Continuing in this manner, we establish the existence of a t_n such that for all ω either sampling has stopped by t_n or

$$(\mu, \rho)[\omega(t_n)] \in \bigcup_{i=1}^{n} A_i$$

Letting $t = t_n$ completes the proof.

Theorem I has three important implications. The first is substantive. The proof was based on the notion that those who fail to find low prices in their initial searches despair of ever finding them and become willing to accept high prices. In other words, persons who continue searching become less finicky as they search more. This seems an important property of search models. For example, in two very different papers, Kasper (1967) and Diamond (1971) presume such behavior and examine its consequences. They do not prove that searchers do behave this way, instead they state that economic theory (diminishing returns) implies they should. Gronau (1971), Gould (1972), and Salop (1973) provid proofs that searchers' standards fall the longer they remain in the market. Their proofs work for different reasons than the one given here. In the Gould and Gronau models, search takes time which is a scarce resource whose real value rises as more of it has been used, in effect, the cost of search increases as search goes on. In the Salop model, searchers' standards fall because searchers first explore the more attractive possibilities. If these prove unsatisfactory, they demand less of what remains.

The other consequences of Theorem I are somewhat more technical. The theorem suggests that computing optimal-search rules and th expected costs of following them is a finite problem, thus, it should b possible to compute the loss from following such ad hoc rules as tho discussed by Telser (1973) instead of following optimal rules. Finally, shows that to prove propositions about $V(\mu, \rho)$ it is sufficient to pro them for all $V_T(\mu, \rho)$. This technique is used repeatedly in the sequ For example, it may be used to prove that property 3 holds in general.

Proposition 3. As the costs of search increase, search decreases.

Proof: Since, according to the optimal-search rule, the search contin whenever $p_i \geq V[h_i(\mu, \rho)] + c$, it will suffice to show that if $V(\mu, \rho$ the value of having information (μ, ρ) when costs are c and $\hat{V}(\mu, \rho$

the value of having the same information when costs are $\hat{c} \geq c$, then $\hat{V}(\mu, \rho) \geq V(\mu, \rho)$. Theorem I implies that this will be true if, for all T, $\hat{V}_T(\mu, \rho) \geq V_T(\mu, \rho)$. But this is clearly true if $T = 0$, that it is true for all T follows by induction from (12b), the definition of V_T.

VI. Reservation Prices

When the distribution of prices is known, optimal-search rules are characterized by a reservation price, there is a price p_R such that the searcher will accept all prices less than or equal to p_R and reject all prices greater than p_R. When the distribution of prices is unknown, acceptable prices change as information changes, so that optimal Bayesian search procedures cannot be characterized by a single reservation price. It is, however, worthwhile to ask whether or not they have a reservation price property, that is, whether for every state of information (μ, ρ) there is a $p_{R(\mu,\rho)}$ such that prices below $p_{R(\mu,\rho)}$ are accepted and those above it are rejected, if the acceptance sets of all customers in a market are characterized by reservation prices, then the demand function of each seller in that market is well behaved—if he raises prices, expected sales will not increase. Raising prices may increase sales if customers' acceptance sets are not characterized by reservation prices.

In general, optimal Bayesian search rules do not have the reservation-price property. Counterexamples are easy to construct. Suppose there are three prices, \$1.00, \$2.00, and \$3.00, and that the cost of search is \$0.01. Prior beliefs admit the possibility of only two distributions of prices. Either all prices are \$3.00 or they are distributed between \$1.00 and \$2.00 in the proportions 99 to 1. A man with these beliefs should accept a price of \$3.00 (as this is a signal that no lower prices are to be had) and reject a quote of \$2.00 (which indicates that the likelihood that a much better price will be obtained on another draw is high).

It is easy to see what makes this counterexample work. Price quotations have value as information. If my beliefs are (μ, ρ) then, the informational value of getting price i is just $V[h_i(\mu, \rho)]$. In the counterexample above, the differences in the information value of prices far exceed the differences in the prices themselves. If differences in the value of price information are less than differences in prices, then optimal-search rules have the reservation-price property.

Proposition 4: If for all $(\mu, \rho) \in \Delta$,

$$|p_i - p_k| \geq |V[h_i(\mu, \rho)] - V[h_k(\mu, \rho)]|, \tag{16}$$

then

$$V[h_i(\mu, \rho)] + c \geq p_i \geq p_k \tag{17}$$

implies

$$V[h_k(\mu, \rho)] + c \geq p_k \tag{18}$$

Note that (17) implies (18) is precisely the reservation-price property.

Proof: Suppose that (16) holds Then, using (17),

$$|p_i - p_k| = p_i - p_k \geq |V[h_i(\mu, \rho)] - V[h_k(\mu, \rho)]|$$
$$\geq V[h_i(\mu, \rho)] - V[h_k(\mu, \rho)]$$

Thus

$$p_k - V[h_k(\mu, \rho)] + c \leq p_i - V[h_i(\mu, \rho)] + c \leq 0$$

This is a useful criterion It is used to prove

Theorem II · Optimal rules for searchers with Dirichlet priors have the reservation-price property

In view of proposition 1 and equation (9), this is about the most general result that could be hoped for in this context The proof which follows is purely formal and conveys little insight It is not, however, hard to see why the theorem should be true Proposition 4 implies that search rules will have the reservation price property whenever observing a price does not convey information about the relative likelihood—as measured by $\lambda(\mu, \rho)$—of observing other prices in the future. The rule for updating the Dirichlet is completely neutral in this sense For all distinct i, j, and k, observing p_i has no effect on λ_j/λ_k

Proof In view of propositions 1 and 4, it will suffice to show that

$$\lambda_i(\mu, \rho) = \mu_i \tag{19}$$

implies that, for all positive integers, s,

$$|p_i - p_k| \geq |V[h_i^s(\mu, \rho)] - V[h_k^s(\mu, \rho)]|, \tag{20}$$

where

$$h_i^s(\mu, \rho) = \left(\frac{\mu_1}{1 + s\rho}, \quad , \frac{\mu_i + s\rho}{1 + s\rho}, \quad , \frac{\mu_n}{1 + s\rho}, \frac{\rho}{1 + s\rho}\right) \tag{21}$$

is (μ, ρ) updated according to (5) after s observations of p_i Note that (16) is just (20) with $s = 1$ Theorem I implies that it will suffice to show, for all t,[13]

$$|p_i - p_k| \geq |V_t(h_i^s) - V_t(h_k^s)|, \tag{22}$$

which can be done by induction Suppose, as we shall throughout the proof, that

$$p_i > p_k \tag{23}$$

Then from (12) and (21),

$$V_0(h_i^s) = \frac{1}{1 + s\rho} \sum_j \mu_j p_j + \frac{s\rho}{1 + s} p_i$$

---

[13] The arguments (μ, ρ) will be suppressed where this can be done without confusion

Thus,

$$V_0(h_i^s) - V_0(h_k^s) = \frac{s\rho}{1 + s\rho}(p_i - p_k) \leq (p_i - p_k),$$

so that (22) holds for $t = 0$. The inductive step that if (22) holds for $t = T - 1$ then it holds for $t = T$ is proved in a series of lemmas below The first is of independent interest It states that if $p_i \geq p_k$, the man who has observed p_i s times expects to incur more total costs, other things equal, than the man who has observed the lower p_k s times

Lemma 1: $p_i \geq p_k$ implies $V_T(h_i^s) \geq V_T(h_k^s)$

Proof: I have already shown that

$$V_0(h_i^s) - V_0(h_k^s) = \frac{s\rho}{1 - s\rho}(p_i - p_k) > 0$$

Suppose the lemma true for $t = T - 1$ Then letting

$$h_j h_i^s = h_j[h_i^s(\mu, \rho)] = h_i^s h_j,$$

$$\begin{aligned} V_T(h_i^s) - V_T(h_k^s) = \frac{1}{1 + \rho} \sum_j \mu_j \{\min\,[p_j, V_{T-1}(h_i^s h_j) + c] \\ - \min\,[p_j, V_{T-1}(h_k^s h_j) + c]\} \\ + \frac{\rho}{1 + \rho}\{\min\,[p_i, V_{T-1}(h_i^{s+1}) + c] \\ - \min\,[p_k, V_{T-1}(h_k^{s+1}) + c]\} \end{aligned} \tag{24}$$

The terms in braces which multiply $1/(1 + \rho)$ are of the form $J(A, B, C) = \min\,(A, B) - \min\,(A, C)$, where $B \geq C$ by the induction hypotheses Thus, they are all nonnegative Consider the term multiplying $\rho/(1 + \rho)$ It is of the form

$$J(A, B, C, D) = \min\,(A, B) - \min\,(C, D), \tag{25}$$

where $A = p_i$, $B = V_{T-1}(h_i^{s+1}) + c$, $C = p_k$, and $D = V_{T-1}(h_k^{s+1}) + c$ Thus, $A \geq C$ and $B \geq D$ by the induction hypotheses. These two inequalities imply that (25) is nonnegative. Since (24) is a weighted average of nonnegative quantities, it is nonnegative

Lemma 2. If (22) holds for $t = T - 1$, then

$$\min\,[p_j, V_{T-1}(h_j h_i^s) + c] - \min\,[p_j, V_{T-1}(h_j h_k^s) + c] \leq p_i - p_k. \tag{26}$$

Proof: The left-hand side of (26) has the form

$$J(A, B, C) = \min\,(A, B) - \min\,(A, C),$$

where $A = p_j$, $B = V_{T-1}(h_j h_i^s) + c$, and $C = V_{T-1}(h_j h_k^s) + c$. Together (22) and lemma 1 imply

$$0 \leq B - C \leq p_i - p_k \tag{27}$$

If $B \leq A$, then $C \leq A$ and $J = B - C$ If $B > A$, either $C \leq A$ and $J = A - C \leq B - C$ or $C > A$ and $J = A - A = 0 \leq B - C$ Thus, $J \leq (B - C) \leq p_i - p_k$

Lemma 3 If (22) holds for $t = T - 1$,

$$\min [p_i, V_{T-1}(h_i^{s+1}) + c] - \min [p_k, V_{T-1}(h_k^{s+1}) + c] \leq p_i - p_k \tag{28}$$

Proof: The left-hand side of (28) is of the same general form as (25), with $A = p_i$, $B = V_{T-1}(h_i^{s+1}) + c$, $C = p_k$, and $D = V_{T-1}(h_k^{s+1}) + c$ Also, by hypothesis, $A \geq C$ and lemma 1 imply $B \geq D$ We want to show that J [illegible] $A - C$ There are four cases to consider

i) $A \leq B$ and $C \leq D$, so that $J = A - C$

ii) $A \leq B$ and $C > D$ This case may be disregarded, since, if it obtains, $B - D \geq A - D > A - C$ or $V_{T-1}(h_i^{s+1}) - V_{T-1}(h_k^{s+1}) > p_i - p_k$, which contradicts the induction hypothesis (22)

iii) $A > B$ and $C > D$, so that $J = (B - D) \leq A - C$ by (22)

iv) $A > B$ and $C \leq D$, so that $J = (B - C) < A - C$, since $A > B$

Lemma 4 If (22) holds for $t = T - 1$, it holds for $t = T$

Proof Use lemmas 2 and 3 to calculate

$$\begin{aligned}
|V_T(h_i^s) - V_T(h_k^s)| &= V_T(h_i^s) - V_T(h_k^s) \\
&= \frac{1}{1 + s\rho} \sum_j \mu_j \{\min [p_j, V_{T-1}(h_j h_i^s) + c] \\
&\qquad - \min [p_j, V_{T-1}(h_j h_k^s) + c]\} \\
&\quad + \frac{s\rho}{1 + s\rho} \{\min [p_i, V_{T-1}(h_i^{s+1}) + c] \\
&\qquad - \min [p_k, V_{T-1}(h_k^{s+1}) + c]\} \\
&\leq \frac{1}{1 + s\rho} \sum_j \mu_j (p_i - p_k) + \frac{s\rho}{1 + s\rho} (p_i - p_k) \\
&= (p_i - p_k) = |p_i - p_k|
\end{aligned}$$

This completes the proof of Theorem II.

Theorem I implies that the reservation price of those searchers who remain in the market must eventually increase to p_n [14]

---

[14] In previous versions of this paper, I conjectured that if the prior were Dirichlet, the reservation price would increase monotonically I am grateful to Emmett Keeler for showing me how to construct a counterexample to this conjecture

VII. Effects of Increasing Uncertainty

The Introduction lists some results on the effects of increased price dispersion when the price distribution is known. It is natural to ask whether these results—that increased price dispersion lowers total expected costs and increases search activity—hold when the distribution is unknown. Since the searcher's knowledge of the price distribution is defined by the parameters (μ, ρ), the question is what effect changes in μ and ρ which represent increased uncertainty have on $V(\mu, \rho)$. To answer this, it is necessary to decide what changes in (μ, ρ) correspond to increased uncertainty. There are two obvious candidates. $\lambda(\mu, \rho)$ is the expected price distribution in the sense that the searcher is just willing to bet that the next price will be p_i at odds $\lambda_i(\mu, \rho)$ to $[1 - \lambda_i(\mu, \rho)]$. However, ρ represents the precision of the searcher's knowledge of the price distribution he faces ($\rho = 0$ is subjective certainty). Thus, $\lambda(\mu, \rho)$ represents what the searcher believes while ρ represents how firmly he believes it. Increased uncertainty could correspond either to an increase in the dispersion of $\lambda(\mu, \rho)$, holding ρ constant, or to an increase in ρ, holding $\lambda(\mu, \rho)$ constant. With what I hope is a pardonable abuse of language, I shall call the first case "increasing perceived uncertainty" and the second case "increasing confidence." This terminology is justified by the fact that increasing perceived uncertainty is the natural analogue of the increases in uncertainty studied when price distributions are assumed known. Both cases are discussed below.

The effects of increased perceived uncertainty, at least for the Dirichlet case, are straightforward.

Theorem III. If the searcher's prior is a Dirichlet $[\lambda(\mu, \rho) = \mu]$, then if $\tilde{\mu}$ is riskier than μ, $V(\tilde{\mu}, \rho) \leq V(\mu, \rho)$.

Proof. Suppose $\tilde{\mu}$ is riskier than μ. It will suffice to prove the theorem if μ and $\tilde{\mu}$ differ by a single mean preserving spread (Rothschild and Stiglitz 1970), so we shall consider four prices $p_1 < p_2 < p_3 < p_4$ such that $\mu_i = \tilde{\mu}_i$ for $i \notin (1, 2, 3, 4)$ and

$$\tilde{\mu}_1 - \mu_1 = \mu_2 - \tilde{\mu}_2 = A > 0 \tag{29}$$

and

$$\mu_3 - \tilde{\mu}_3 = \tilde{\mu}_4 - \mu_4 = B > 0, \tag{30}$$

where

$$A(p_2 - p_1) + B(p_3 - p_4) = 0 \tag{31}$$

I shall prove the theorem by showing

$$V_T(\tilde{\mu}, \rho) \leq V_T(\mu, \rho) \tag{32}$$

for all T by induction. That (32) holds for $T = 0$ follows from (31) and (12). In an obvious notation, let $\tilde{h}_j = h_j(\tilde{\mu}, \rho)$.

Let R satisfy

$$V_T(\mu, \rho) = \sum_{1}^{R} \mu_j p_j + \sum_{R+1}^{N} \mu_j[V_{T-1}(h_j) + c]$$

Theorem II guarantees the existence of such an R. Note that

$$\begin{aligned} V_T(\tilde{\mu}, \rho) &= \sum \tilde{\mu}_j \min\,[p_j, V_{T-1}(\tilde{h}_j) + c] \\ &\leq \sum_{1}^{R} \tilde{\mu}_j p_j + \sum_{R+1}^{n} \tilde{\mu}[V_{T-1}(\tilde{h}_j) + c] \\ &\leq \sum_{1}^{R} \tilde{\mu}_j p_j + \sum_{R+1}^{n} \tilde{\mu}[V_{T-1}(h_j) + c] \end{aligned}$$

(the last inequality follows from the induction hypothesis) Let

$$H(\text{[illegible]}\tilde{\mu}) = \sum_{1}^{R} (\mu_j - \tilde{\mu}_j)p_j + \sum_{R+1}^{n} (\mu_j - \tilde{\mu}_j)[V_{T-1}(h_j) + c] \quad (33)$$

Then it will suffice to show that $H(\mu, \tilde{\mu}) \geq 0$

There are five cases to consider, depending on the relationship of p_R to p_1, p_2, p_3, and p_4

1) $p_R \geq p_4$ In this case,

$$\begin{aligned} H(\mu, \tilde{\mu}) &= (\mu_1 - \tilde{\mu}_1)p_1 + (\mu_2 - \tilde{\mu}_2)p_2 + (\mu_3 - \tilde{\mu}_3)p_3 + (\mu_4 - \tilde{\mu}_4)p_4 \\ &= A(p_2 - p_1) + B(p_3 - p_4) = 0 \end{aligned}$$

2) $p_4 > p_R > p_3$ In this case,

$$\begin{aligned} H(\mu, \tilde{\mu}) &= A(p_2 - p_1) + B\{p_3 - [V_{T-1}(h_4) + c]\} \\ &\geq A(p_2 - p_1) + B(p_3 - p_4) = 0 \end{aligned}$$

3) $p_3 > p_R > p_2$ In this case,

$$\begin{aligned} H(\mu, \tilde{\mu}) &= A(p_2 - p_1) + B[V_{T-1}(h_3) - V_{T-1}(h_4)] \\ &\geq A(p_2 - p_1) + B(p_3 - p_4) = 0 \end{aligned}$$

The second inequality follows from (22)

4) $p_2 > p_R \geq p_1$ In this case,

$$\begin{aligned} H(\mu, \tilde{\mu}) &= A[V_{T-1}(h_2) + c - p_1] + B[V_{T-1}(h_3) - V_{T-1}(h_4)] \\ &\geq A[V_{T-1}(h_2) - V_{T-1}(h_1)] + B[V_{T-1}(h_3) - V_{T-1}(h_4)] \end{aligned} \quad (34)$$

It is shown below in lemma 5 that

$$\frac{V_{T-1}(h_2) - V_{T-1}(h_1)}{p_2 - p_1} \geq \frac{V_{T-1}(h_4) - V_{T-1}(h_3)}{p_4 - p_3}, \quad (35)$$

so that

$$A[V_{T-1}(h_2) - V_{T-1}(h_1)] \geq \frac{A(p_2 - p_1)}{p_4 - p_3}[V_{T-1}(h_4) - V_{T-1}(h_3)]$$

Combining (34) and (35), we obtain

$$H(\mu, \tilde{\mu}) \geq \frac{V_{T-1}(h_4) - V_{T-1}(h_3)}{p_4 - p_3}[A(p_2 - p_1) + B(p_4 - p_3)] = 0.$$

5) $p_1 \geq p_R$. In this case, $H(\mu, \tilde{\mu})$ is just equal to the last expression in (34), and (35) can be used to show that $H(\mu, \tilde{\mu}) \geq 0$.

To complete the proof we only need prove

Lemma 5: Inequality (35) holds for all T.

Proof: If $T = 0$, both sides of (35) are equal to $[\rho/(1 + \rho)]$ Suppose (35) holds for $t = T - 1$. Let R_k satisfy

$$V_T(h_k) = \sum_{1}^{R_k} \mu_j p_j + \sum_{R_k+1}^{n} \mu_j[V_{T-1}(h_{jk}) + c]$$

It follows from lemma 1 that $R_1 \leq R_2 \leq R_3 \leq R_4$, so that

$$V_T(h_2) - V_T(h_1) \leq \sum_{R_3+1}^{n} \mu_j[V_{T-1}(h_{2j}) - V_{T-1}(h_{1j})] \qquad (36)$$

while

$$V_T(h_4) - V_T(h_3) \leq \sum_{R_3+1}^{n} u_j[V_{T-1}(h_{4j}) - V_{T-1}(h_{3j})] \qquad (37)$$

The induction hypothesis implies

$$0 \leq \sum_{R_3+1}^{n} \mu_j \left[\frac{V_{T-1}(h_{2j}) - V_{T-1}(h_{1j})}{p_2 - p_1} - \frac{V_{T-1}(h_{4j}) - V_{T-1}(h_{3j})}{p_4 - p_3}\right]$$
$$\leq \frac{V_{T-1}(h_2) - V_{T-1}(h_1)}{p_2 - p_1} - \frac{V_{T-1}(h_4) - V_{T-1}(h_3)}{p_4 - p_3}. \qquad (38)$$

The last inequality follows from (36) and (37)

An immediate consequence of Theorem III is an analogue of property 4 that increased price dispersion leads to increased search

Corollary: If the searcher's prior is Dirichlet, then increasing perceived uncertainty lowers the reservation price

The proof, which is obvious, is omitted

The effects of increased confidence are less clear. I suspect that if the prior is Dirichlet, then in most cases $V(\mu, \rho)$ decreases as ρ decreases However, I have not been able to prove that this is always true or to construct an example where $V(\mu, \rho)$ increases when ρ decreases [15]

---

[15] **Earlier versions of this paper contained what I thought was an example of $V(\mu, \rho)$ increasing when ρ decreased I am grateful to Mr. Richard Ressler for showing me that this example was mistaken Ressler's senior thesis (Princeton University, 1974) contains the details of this and other calculations of the effects of various parameter changes on optimal-search strategies.**

VIII. Conclusion

The most important results of this paper, Theorems II and III, state that, for what proposition 1 suggests is a quite significant example, the qualitative behavior of persons searching optimally from unknown distributions is the same as that of persons searching optimally from known distributions Since it is easy to construct examples for which this is not true, it is natural to ask just how general these results are This problem is best stated by focusing on a very general and abstract formulation of the problem of optimal search. Let K be a compact[16] set of the real line and let $\mathscr{M}(K)$ be the set of all probability measures on K. The quantity $\mathscr{M}(K)$ is a compact separable metric space, so that we can, in the standard way, define probability measures on $\mathscr{M}(K)$[17] Such a probability measure, say $\mathscr{P}$, represents a searcher's a priori beliefs about the price distribution he faces (which is some probability distribution in $\mathscr{M}(K)$) If a price p is observed, then as long as the conditional distributions (on $\mathscr{M}(K)$ given p) can be calculated, Bayes's rule may be used to update his beliefs from $\mathscr{P}$ to, say, $\mathscr{P}'$ As in Sections II and IV above, the optimal Bayesian strategy can be devised The measure $\mathscr{P}$ completely determines the optimal-search rule The question at hand is, For what class of $\mathscr{P}$ do analogues of Theorems II and III hold?

Each $\mathscr{P}$ has two rather distinct aspects The first is $S(\mathscr{P})$, the support of $\mathscr{P}$, which is roughly the set of price distributions which the searcher whose beliefs are described by $\mathscr{P}$ judges to be conceivable This paper focuses on the case where $S(\mathscr{P})$ contains only multinomial distributions The second aspect of $\mathscr{P}$ is the updating rule Theorems II and III require that the updating rule be of a very special kind—if p_i is observed, then the (subjective) likelihood of observing p_i in the future is increased while the likelihood of observing all other prices is decreased (and by the same proportion)

Recent work of Ferguson (1973)[18] suggests that it is only the restrictions on the updating rule and not those on $S(\mathscr{P})$ that are required for Theorems II and III Ferguson introduces a class of distributions whose support is essentially all of $\mathscr{M}(K)$, I conjecture that analogues of Theorems I, II, and III hold for these distributions However, these probability measures imply very special updating rules If we ignore many technicalities, Ferguson's work may be summarized as follows Let α be a finite measure on K and let $B = (B_1, B_2, \quad, B_n)$ be a measurable partition of K (that is, the B_i are measurable and $\bigcup_i B_i = K$ and $B_i \cap B_j = \phi$ if $i \neq j$) Consider the probability that the next observed price belongs to $B_1, B_2,$ or B_n This is a multinomial distribution described by the numbers $(P(B_1), \quad , P(B_n)) \in \Delta$ Each such multinomial distribution belongs to $\mathscr{M}(K)$ and is thus assigned probability by $\mathscr{P}$ The distribution $\mathscr{P}_\alpha$ is a Ferguson distribution with parameter α if the distribution of $(P(B_1), P(B_2), \quad , P(B_n))$ is Dirichlet with parameter $(\alpha(B_1), \alpha(B_2), \quad , \alpha(B_n))$ Ferguson shows (1973, pp. 215–16) that the support of $\mathscr{P}_\alpha$ is the set of all measures $\mu \in \mathscr{M}(K)$ which are absolutely continuous with respect to α. By choosing α appropriately, $S(\mathscr{P}_\alpha)$ can be made very large

---

[16] I continue to insist that price distributions be bounded No meaningful generality and much complication ensues if this assumption is abandoned

[17] See Parthasarthy (1967, pp 39 ff) for details

[18] Blackwell and MacQueen (1973) show how Ferguson distributions can be understood in terms of the urn model of Section III In terms of that model, we can phrase our question as follows Do our results hold when the composition of the urn is altered (upon observation of price p_i) in some other way than by adding another slip with p_i written on it?

The rule for updating $\mathscr{P}_\alpha$ is very simple If $\mathscr{P}_\alpha$ is a Ferguson distribution with parameter α, after price p is observed, it is updated to a Ferguson distribution with parameter

$$h_p(\alpha) = \alpha + \delta_p, \tag{39}$$

where δ_p is the probability measure concentrated on the point p This formula is an exact analogue of the rule given in equation (10) above for updating Dirichlet priors

Proving analogues of Theorems I, II, and III for Ferguson distributions whose parameters have compact support should involve only technical problems [19] If these could be solved, it would show that the results of this paper are quite general in the sense that they apply to searchers whose beliefs are not restricted in any way beyond boundedness However, they are quite special in that they apply only to people who revise their beliefs in a very special way Formula (39) implies that information is strictly local or, to put the matter more colloquially, that a miss is as good as a mile My future beliefs about the likelihood of observing p are affected in exactly the same way by observing $p + 10^{-10}$ as by observing $p + 10^{10}$ I would like to know whether the conclusions of Theorems II and III hold when information is not so completely local At present there seems no way to answer this question with generality I know of no class of $\mathscr{P} \in \mathscr{M}(K)$ which has both large support and tractable updating rules other than the Ferguson distributions

It is possible to examine the property of $\mathscr{P}$'s with rather restricted support by examining parametric families of distributions with tractable updating rules For example, it is a simple consequence of DeGroot's (1968) work that optimal-search rules from normal distributions with unknown means and known variance have the reservation-price property when the prior distribution is also normal

Perhaps the strongest possible result along the lines of Theorem II [20] that could be hoped for is the following

---

[19] It should be possible to push through proofs very similar to those in the text once a topology is found in which proposition 2 holds

[20] I am not certain how propositions about the effects of increasing dispersion are best stated in this context.

Conjecture: Optimal-search rules from exponential families of distributions have the reservation-price property if the unknown parameters have conjugate prior distributions.

Establishing this would significantly expand the applicability of the results of this paper Still, I think enough has been done to establish that economists can without great loss assume that the qualitative properties of demand functions which arise from optimal search from unknown distributions are the same as those which arise from optimal search from known distributions

References

Billingsley, P *Convergence of Probability Measures* New York Wiley, 1968

Blackwell, D, and MacQueen, J B "Ferguson Distributions via Polya Urn Schem[illegible]" *Ann. Statis* 1 (1973) 353–55

Breiman, L[illegible]"Stopping-Rule Problems" In *Applied Combinatorial Mathematics*, edited by E F Beckenbach Berkeley Univ California Press, 1964

Buckley, J E "Wage Dispersion in Metropolitan Areas" *Monthly Labor Rev* 92, no 9 (September 1969) 24–29

Chow, Y S, Robbins, H, and Siegmund, D *Great Expectations The Theory of Optimal Stopping* Boston Houghton Mifflin, 1971

David, P A "Fortune, Risk, and the Micro-Economics of Migration" In *Nations and Households in Economic Growth Essays in Honor of Moses Abramovitz*, edited by P A David and M W Reder New York Academic Press, 1973

De Groot, M H "Some Problems of Optimal Stopping" *J Roy Statis Soc B* 30 (March 1968) 108–22

——— *Optimal Statistical Decisions* New York McGraw-Hill, 1970

Diamond, P A "A Model of Price Adjustment" *J Econ Theory* 3 (June 1971) 156–68

Diamond, P A, and Stiglitz, J E *Increases in Risk and in Risk Aversion J Econ Theory* (1974), in press

Ferguson, T S "A Bayesian Analysis of Some Nonparametric Problems" *Ann Statis* 1 (1973) 209–30

Gastwirth, J L "On Probabilistic Models of Consumer Search for Information." Unpublished paper, 1971

Gould, J P "Information and Consumer Behavior Theorv of Optimal Sequential Searching Policies." Unpublished paper, 1972

Gronau, R "Information and Frictional Unemployment" *A E R* 61 (June 1971)· 290–301

Kasper, H "The Asking Price of Labor and the Duration of Unemployment." *Rev Econ and Statis* 49 (May 1967) 165–73.

Kohn, M G, and Shavell, S "Optimal Adaptive Search" *J Econ Theory* (1974), in press

McCall, J J. "The Economics of Information and Optimal Stopping Rules." *J. Bus.* 38 (July 1965) 300–317.

———. "Economics of Information and Job Search" *Q J.E* 84 (February 1970) 113–26.

Nelson, P. "Information and Consumer Behavior." *J.P E.* 78 (March/April 1970) 311–29.

Parthasarthy, K R *Probability Measures on Metric Spaces* New York Academi Press, 1967.

Rothschild, M. "Models of Market Organization with Imperfect Information." *J.P.E.* 81 (November/December 1973): 1283–1308.

Rothschild, M., and Stiglitz, J. E. "Increasing Risk. I. A Definition" *J. Econ. Theory* 2 (September 1970): 225–43

Salop, S. "Systematic Job Search and Unemployment" *Rev. Econ Studies* 40 (April 1973) 191–202

Stigler, G. J "The Economics of Information." *J P E* 69 (June 1961) 213–25.

———. "Information in the Labor Market." *J P.E.* 70, no 5, suppl (October 1962) · 94–105.

Telser, L G "Searching for the Lowest Price" *A.E.R Proc* 63 (1973) 41–49.

Whipple, D "A Generalized Theory of Job Search" *J P E* 81 (September/October 1973). 1170–88

Yahav, J. "On Optimal Stopping" *Ann Math Statis* 37 (1966) 30–35

Karl Marx on the Economic Role of Science

Nathan Rosenberg

University of Wisconsin

> It is not the articles made, but how they are made, and by what instruments, that enables us to distinguish different economical epochs [MARX 1906, p 200]

This paper examines Marx's treatment of rising resource productivity and technological change under capitalism Little attention has been given to Marx's view of the role which science plays in these processes. It is obvious that Marx (and Engels) attach the greatest importance to the development of modern science, but the way in which scientific progress meshes with the rest of the Marxian system has not been fully understood The paper analyzes Marx's treatment of the factors which account for the growth of scientific knowledge as well as capitalist society's changing capacity to incorporate this knowledge into the productive process

The purpose of this paper is to examine certain aspects of Marx's treatment of rising resource productivity and technological change under capitalism Many of the most interesting aspects of Marx's treatment of technological change have been ignored, perhaps because of the strong polemical orientation which readers from all shades of the political spectrum seem to bring to their reading of Marx As a result, much has been written about the impact of the machine upon the worker and his family, the phenomenon of alienation, the relationship between technological change, real wages, employment, etc At the same time, a great deal of what Marx had to say concerning some 300 years of European capitalist development has received relatively little attention. This applies to his views dealing with the complex interrelations between science, technology, and economic development

It is a well-known feature of the Marxian analysis of capitalism that Marx views the system as bringing about unprecedented increases in

The author is grateful to Professors Stanley Engerman and Eugene Smolensky for critical comments on an earlier draft

human productivity and in man's mastery over nature. Marx and Engels told their readers, in *The Communist Manifesto*, that "the bourgeoisie, during its rule of scarce one hundred years, has created more massive and more colossal productive forces than have all preceding generations together. Subjection of Nature's forces to man, machinery, application of chemistry to industry and agriculture, steam-navigation, railways, electric telegraphs, clearing of whole continents for cultivation, canalisation of rivers, whole populations conjured out of the ground—what earlier century had even a presentiment that such productive forces slumbered in the lap of social labour?" (Marx and Engels 1951, 1:37). No single question, therefore, would seem to be more important to the whole Marxian analysis of capitalist development than the question: Why is capitalism such an immensely productive system by comparison with all earlier forms of economic organization? The question, obviously, has been put before, and certain portions of Marx's answer are in fact abundantly plain. In particular, the social and economic structure of capitalism is one which creates enormous incentives for the generation of technological change. Marx and Engels insist that the bourgeoisie is unique as a ruling class because, unlike all earlier ruling classes whose economic interests were indissolubly linked to the maintenance of the status quo, the very essence of bourgeois rule is technological dynamism.[1] Capitalism generates unique incentives for the introduction of new, cost-reducing technologies.

The question which I am particularly interested in examining is the role which is played, within the Marxian framework, by science and scientific progress in the dynamic growth of capitalism. For surely the growth in resource productivity can never have been *solely* a function of the development of capitalist institutions. It is easy to see the existence of such institutions as a necessary condition but hardly as a sufficient condition for such growth. Surely the technological vitality of an emergent capitalism was closely linked up with the state of scientific knowledge and with industry's capacity to exploit such knowledge.

Marx's (and Engels's) position, briefly stated, is to affirm that science is, indeed, a fundamental factor accounting for the growth in resource productivity and man's enlarged capacity to manipulate his natural environment for the attainment of human purposes. However, the statement requires two immediate and highly significant qualifications, which will constitute our major concern in this paper: (1) science does not, according to Marx, function in history as an independent variable, and (2) science has come to play a critical role as a systematic contributor to increasing productivity only at a very recent (from Marx's perspective) point in history. The ability of science to perform this role had necessarily to await the fulfillment of certain objective conditions. What these conditions were has not been understood adequately.

---

[1] "The bourgeoisie cannot exist without constantly revolutionising the instruments of production, and thereby the relations of production, and with them the whole relations of society. Conservation of the old modes of production in unaltered form, was, on the contrary, the first condition of existence for all earlier industrial classes" (Marx and Engels 1951, 1:36).

I

Marx's treatment of scientific progress is consistent with his broader historical materialism. Just as the economic sphere and the requirements of the productive process shape man's political and social institutions, so do they also shape his scientific activity at all stages of history. Science does not grow or develop in response to forces internal to science or the scientific community. It is not an autonomous sphere of human activity. Rather, science needs to be understood as a social activity which is responsive to economic forces. It is man's changing needs as they become articulated in the sphere of production which determine the direction of scientific progress. Indeed, this is generally true of all human problem-solving activity, of which science is a part. As Marx states in the introduction to his *Critique of Political Economy*: "Mankind always takes up only such problems as it can solve, since, looking at the matter more closely, we will always find the problem itself arises only when the material conditions necessary for its solution already exist or are at least in the process of formulation" (Marx 1904, pp. 12–13).

Marx views specific scientific disciplines as developing in response to problems arising in the sphere of production. The materialistic conception of history and society involves the rejection of the notion that man's intellectual pursuits can be accorded a status independent of material concerns. It emphasizes the necessity of systematically relating the realm of thinking and ideas to man's material concerns. Thus, the scientific enterprise itself needs to be examined in that perspective. "Feuerbach speaks in particular of the perception of natural science, he mentions secrets which are disclosed only to the eye of the physicist and chemist; but where would natural science be without industry and commerce? Even this 'pure' natural science is provided with an aim, as with its material, only through trade and industry, through the sensuous activity of men" (Marx and Engels 1947, p. 36). Egyptian astronomy had developed out of the compelling need to predict the rise and fall of the Nile, upon which Egyptian agriculture was so vitally dependent (Marx 1906, p. 564, n. 1). The increasing (if still "sporadic") resort to machinery in the seventeenth century was, says Marx, "of the greatest importance, because it supplied the great mathematicians of that time with a practical basis and stimulant to the creation of the science of mechanics."[2] The difficulties encountered with gearing as waterpower was being harnessed to larger millstones was "one of the circumstances that led to a more accurate investigation of the laws of friction "[3]

---

[2] Marx 1906, pp. 382–83. Engels states: "Like all other sciences, mathematics arose out of the *needs* of men; from the measurement of land and of the content of vessels, from the computation of time and mechanics" (Engels 1939, p. 46; emphasis Engels's. Cf. Marx 1906, p. 564)

These themes are repeated by Engels, who asserts that "from the very beginning the origin and development of the sciences has been determined by production."[4] In accounting for the rise of science during the Renaissance, his first explanation again drew upon the requirements of industry

> If, after the dark night of the Middle Ages was over, the sciences suddenly arose anew with undreamt-of force, developing at a miraculous rate, once again we owe this miracle to production In the first place, following the crusades, industry developed enormously and brought to light a quantity of new mechanical (weaving, clock-making, milling), chemical (dyeing, metallurgy, alcohol), and physical (spectacles) facts, and this not only gave enormous material for observation, but also itself provided quite other means for experimenting than previously existed, and allowed the construction of *new* instruments, it can be said that really systematic experimental science now became possible for the first time [5]

Moreover, in a letter written in 1895, Engels stated "If, as you say, technique largely depends on the state of science, science depends far more still on the *state* and the *requirements* of technique If society has a technical need, that helps science forward more than ten universities The whole of hydrostatics (Torricelli, etc) was called forth by the necessity for regulating the mountain streams of Italy in the sixteenth and seventeenth centuries We have only known about electricity since its technical applicability was discovered" (Marx and Engels 1951, 2 457, Letter from Engels to H Starkenburg, January 25, 1895, emphasis Engels's)

---

[3] Marx 1906, p 411 He adds "In the same way the irregularity caused by the motive power in mills that were put in motion by pushing and pulling a lever, led to the theory, and the application, of the flywheel, which afterwards plays so important a part in Modern Industry In this way, during the manufacturing period, were developed the first scientific and technical elements of Modern Mechanical Industry "

[4] Engels 1954, p 247 Earlier in the paragraph, he had stated "The successive development of the separate branches of natural science should be studied First of all, astronomy, which, if only on account of the seasons, was absolutely indispensable for pastoral and agricultural peoples. Astronomy can only develop with the aid of mathematics Hence this also had to be tackled Further, at a certain stage of agriculture and in certain regions (raising of water for irrigation in Egypt), and especially with the origin of towns, big building structures and the development of handicrafts, mechanics also arose This was soon needed also for navigation and war Moreover, it requires the aid of mathematic and so promoted the latter's development "

[5] Ibid., p 248. The editor of Engels's unfinished manuscript points out that Enge had written in the margin of the manuscript opposite this paragraph· "Hitherto, wh has been boasted of is what production owes to science, but science owes infinitely mo to production "

This statement is probably the most explicit and direct assertion in the writings of Marx and Engels that factors affecting the demand for science are overwhelmingly more important than factors affecting its supply. Scientific knowledge is acquired when a social need for that knowledge has been established Science is, however, not an initiating force in the dynamics of social change Developments in this sphere are a response to forces originating elsewhere. Thus, Marx and Engels appear to be presenting a purely demand-determined explanation of the social role of science. Scientific enterprise supplies that which industry demands, and therefore the changing direction of the thrust of science needs to be understood in terms of the changing requirements of industry

II

In this section I will argue that, while the demand-oriented component of the argument just presented is indeed a major part of the Marxian view, there are also vital but less conspicuous elements in Marx's argument which have been ignored Without these additional and more neglected elements one cannot explain a central thesis which emerges out of Marx's view namely, that it is only at a particular time in human history that science is enlisted in a crucial way in the productive process It is only at a very recent point in history, Marx argues, that the marriage of science and industry occurs Moreover, this marriage does *not* coincide with the historical emergence of capitalism In fact, Marx is quite explicit that the union of science and industry comes only centuries after the arrival of modern capitalism and the emergence of sophisticated bodies of theoretical science If arguments based upon the existence of capitalist incentives and demand forces generally were a sufficient explanation, the full-scale industrial exploitation of science would have come at a much earlier stage in Western history But it did not Why?

Stripped to its essentials, Marx's answer is that the handicraft and manufacturing stages of production lacked the technological basis which would *permit* the application of scientific knowledge to the solution of problems of industrial production.[6] This essential technological basis emerged only with modern industry The immense and growing productivity of nineteenth century British industry was really, in Marx's view, the resultant of three converging sets of forces· (1) the unique incentive system and capacity for accumulation provided by capitalist institutions, (2) the availability of bodies of scientific knowledge[7] which were directly relevant for problem-solving activities in industry, and (3) a technology possessing certain special characteristics It is this last category which is least understood and to which we therefore now turn

---

[6] Since the subsequent discussion turns directly upon the Marxian periodization scheme, it is important to remind the reader of the meaning which Marx attaches to the terms "handicraft," "manufacture," and "modern industry" Engels expressed Marx's meanings succinctly as follows "We divide the history of industrial production since the Middle Ages into three periods· (1) handicraft, small master craftsmen with a few journeymen and apprentices, where each laborer produces the complete article; (2) manufacture, where greater numbers of workmen, grouped in one large establishment, produce the complete article on the principle of division of labor, each workman performing only one partial operation, so that the product is complete only after having passed successively through the hands of all; (3) modern industry, where the product is produced by machinery driven by power, and where the work of the laborer is limited to superintending and correcting the performances of the mechanical agent" (Engels 1910, pp 12–13).

Historically, capitalist relationships were introduced in an unobtrusive way, by the mere quantitative expansion in the number of wage-laborers employed by an individual owner of capital (Marx 1906, p 367) The independent handicraftsman, operating with a few journeymen and apprentices, gradually shifted into the role of a capitalist as his relationship with these men assumed the form of a permanent system of wage payments and as the number of such laborers increased [8] The system of manufacture, therefore, while introducing social relationships drastically different from the handicraft system of the medieval guilds which preceded it,[9] initially employed the same technology [10]

From Marx's mid-nineteenth-century vantage point, the system of manufacture had actually been the dominant one throughout most of the history of capitalism—from "roughly speaking the middle of the 16th to the last third of the 18th century" (Marx 1906, p 369, see also p 787)

---

[7] Actually, Marx's use of the term "science" was sufficiently broad that it included bodies of systematized knowledge far beyond what we ordinarily mean when we speak today of pure or even applied science—e.g , engineering and machine building It was not a term which he attempted to use with precision In *Theories of Surplus Value*, for instance, he refers to science simply as "the product of mental labour" (Marx 1963, pt 1, p 353)

[8] "With regard to the mode of production itself, manufacture, in its strict meaning, is hardly to be distinguished, in its earliest stages, from the handicraft trades of the guilds, otherwise than by the greater number of workmen simultaneously employed by one and the same individual capital The workshop of the medieval master handicraftsman is simply enlarged" (Marx 1906, p 353 Cf Marx and Engels 1947, pp 12–13)

[9] "With manufacture was given simultaneously a changed relationship between worker and employer In the guilds the patriarchal relationship between journeyman and master maintained itself, in manufacture its place was taken by the monetary relation between worker and capitalist—a relationship which in the countryside and in small towns retained a patriarchal tinge, but in the larger, the real manufacturing towns, quite early lost almost all patriarchal complexion" (Marx and Engels 1947, p 52)

[10] Machinery had sometimes been employed in earlier periods, but Marx clearly regarded these instances as exceptional "Early in the manufacturing period the principle of lessening the necessary labour-time in the production of commodities, was accepted and formulated and the use of machines, especially for certain simple first processes that have to be conducted on a very large scale, and with the application of great force, sprang up here and there Thus, at an early period in paper manufacture, the tearing up of the rags was done by paper mills, and in metal works, the pounding of the ores was effected by stamping mills The Roman Empire had handed down the elementary form of all machinery in the water-wheel" (Marx 1906, p 382) In a footnote Marx makes the extremely interesting observation that "the whole history of the development of machinery can be traced in the history of the corn mill" (ibid , p 382, n. 3)

Manufacture involved a significant regrouping of workers and a redefinition of the responsibilities of each. Whereas a medieval handicraftsman would himself perform a succession of operations upon a product, the manufacturing system divided up the operation into a succession of steps, each one of which was allocated to a separate workman [11]

The essence of the manufacturing system, therefore, is a growing specialization on the part of the individual worker. While this in turn has psychological and social consequences of the greatest importance for the worker with which Marx was very much concerned,[12] it continued to share with the earlier handicraft system an essential feature That is to say, although the product now passed through a succession of hands, and although this reorganization raised the productivity of labor, it nevertheless perpetuated the industrial system's reliance upon human skills and capacities [13] Whereas the critical skill was formerly that of the guild craftsman, it is now the unremitting repetition of a narrowly defined activity on the part of the detail laborer More precisely, the productive process now pressed against the constraints imposed by the limited strength, speed, precision, and, indeed, the limited number of limbs, of the human animal

So long as the worker continues to occupy strategic places in the productive process, that process is limited by all of his human frailties And, of course, the individual capitalist is, in many ways, continually pressing the worker against those limits But the point which Marx is making here is of much broader significance The application of science to the productive process involves dealing with impersonal laws of nature and freeing itself from all dependence upon the organic It involves calculations concerning the behavior of natural phenomena It involves the exploitation of reliable physical relationships which have been established by scientific disciplines It involves a degree of predictability of a purely objective sort, from which the uncertainties and subjectivities of human behavior have been systematically excluded Science, in short, can only incorporate its findings in impersonal machinery. It cannot be incorporated in human beings with their individual volitions, idiosyncracies, and refractory temperaments The manufacturing period shared with the earlier handicraft system the essential feature that it was a tool-using economy where the tools were subject to human manipulation and guidance It is this element of human control, the continued reliance upon the limited range of activities of the human hand, and not the nature of the power source, Marx insists, which is decisive [illegible] distinguishing a machine from a tool

---

[11] "The needlemaker of the Nuremberg Guild was the cornerstone on which the English needle manufacture was raised But while in Nuremberg that single artificer performed a series of perhaps 20 operations one after another, in England it was not long before there were 20 needlemakers side by side, each performing one alone of those 20 operations, and in consequence of further experience, each of those 20 operations was again split up, isolated, and made the exclusive function of a separate workman" (ibid , pp 370–71)

[12] 'While simple co-operation leaves the mode of working by the individual for the most part unchanged, manufacture thoroughly revolutionises it, and seizes labour-power by its very roots It converts the labourer into a crippled monstrosity, by forcing his detail dexterity at the expense of a world of productive capabilities and instincts, just as in the States of La Plata they butcher a whole beast for the sake of his hide or his tallow" (ibid , p 396)

[13] "For a proper understanding of the division of labour in manufacture, it is essential that the following points be firmly grasped First, the decomposition of a process of production into its various successive steps coincides, here, strictly with the resolution of a handicraft into its successive manual operations Whether complex or simple, each operation has to be done by hand, retains the character of a handicraft, and is therefore dependent on the strength, skill, quickness, and sureness, of the individual workman in handling his tools The handicraft continues to be the basis This narrow technical basis excludes a really scientific analysis of any definite process of industrial production, since it is still a condition that each detail process gone through by the product must be capable of being done by hand and of forming, in its way, a separate handicraft. It is just because handicraft skill continues, in this way, to be the foundation of the process of production that each workman becomes exclusively assigned to a partial function, and that for the rest of his life, his labour-power is turned into the organ of this detail function" (ibid , pp. 371–72).

> The machine proper is a mechanism that, after being set in motion, performs with its tools the same operations that were formerly sone by the workman with similar tools Whether the motive power is derived from man, or from some other machine, makes no difference in this respect From the moment that the tool proper is taken from man, and fitted into a mechanism, a machine takes the place of a mere implement The difference strikes one at once, even in those cases where man himself continues to be the prime mover The number of implements that he himself can use simultaneously, is limited by the number of his own natural instruments of production, by the number of his bodily organs The number of tools that a machine can bring into play simultaneously, is from the very first emancipated from the organic limits that hedge in the tools of a handicraftsman [14]

III

What, then, is the distinctive technological feature of modern industry? It is that, for the first time, the design of the productive process is carried out on a basis where the characteristics of the worker and his physical endowment are no longer central to the organization and arrangement of capital. Rather, capital is being designed in accordance with a completely different logic, a logic which explicitly incorporates principles of science and engineering.[15] The subjectivity of a technology adapted, out of necessity, to the capacities (or, better, the debilities) of the worker is rejected in favor of the objectivity of machinery which has been designed in accordance with its own laws and the laws of science

---

[14] Ibid , p. 408, see also p 410 In his early work, *The Poverty of Philosophy*, Marx had stated "The machine is a unification of the instruments of labour, and by no means a combination of different operations for the worker himself. 'When, by the division of labour, each particular operation has been simplified to the use of a single instrument, the linking-up of all these instruments, set in motion by a single engine, constitutes—a machine ' (Babbage, *Traité sur l'Economie des Machines*, etc , Paris 1833). Simple tools, accumulation of tools, composite tools, setting in motion of a composite tool by a single hand engine, by men, setting in motion of these instruments by natural forces, machines, system of machines having one motor, system of machines having one automatic motor—this is the progress of machinery" (Marx, n d , pp 132–33 This book was first published in 1847).

> In Manufacture it is the workmen who, with their manual implements, must, either singly or in groups, carry on each particular detail process If, on the one hand, the workman becomes adapted to the process, on the other, the process was previously made suitable to the workman This subjective principle of the division of labour no longer exists in production by machinery Here, the process as a whole is examined objectively, in itself, that is to say, without regard to the question of its execution by human hands, it is analysed into its constituent phases, and the problem, how to execute each detail process, and bind them all into a whole, is solved by the aid of machines, chemistry, etc [16]

The shift from the hand-operated to the machine-operated process is a momentous one, for the simple reason that machine processes are susceptible to continuous and indefinite improvement, whereas hand processes are not [17] The factory system makes possible the virtual routinization of productivity improvement [18] By breaking down the productive process into objectively identifiable component parts, it creates a structure of activities which is readily amenable to rigorous analysis. "The principle, carried out in the factory system, of analysing the process of production into its constituent phases, and of solving the problems thus proposed by the application of mechanics, of chemistry, and of the whole range of the natural sciences, becomes the determining principle everywhere."[19] Thus, historical development has brought technology to a point where it has become, for the first time, an object of scientific analysis and improvement.

---

[15] There is an important learning experience at the technological level before this can be done well "It is only after considerable development of the science of mechanics, and accumulated practical experience, that the form of a machine becomes settled entirely in accordance with mechanical principles, and emancipated from the traditional form of the tool that gave rise to it" (Marx 1906, p 418, n 1) A typical aspect of the innovation process, therefore, is that machines go through a substantial process of modification after their first introduction (see ibid , p 442)

[16] Ibid , pp 414–15 Later, Marx adds "The implements of labour, in the form of machinery, necessitate the substitution of natural forces for human force, and the conscious application of science, instead of rule of thumb In Manufacture, the organization of the social labour-process is purely subjective, it is a combination of detail labourers, in its machinery system, Modern Industry has a productive organism that is purely objective, in which the labourer becomes a mere appendage to an already existing material condition of production" (p 421)

[17] "As soon as a machine executes, without man's help, all the movements requisite to elaborate the raw material, needing only attendance from him, we have an automatic system of machinery, and one that is susceptible of constant improvement in its details" (ibid , p. 416)

[18] In a valuable article, "Karl Marx and the Industrial Revolution," Paul Sweezy argues that many of the important differences between Marx and his classical predecessors reduced to the fact that the classical economists "took as their model an economy based on manufacture, which is an essentially conservative and change-resistant economic order, while Marx, recognizing and making full allowance for the profound transformation effected by the industrial revolution, took as his model an economy based on modern machine industry" (Sweezy 1968, p 115)

> A characteristic feature is, that, even down into the eighteenth century, the different trades were called "mysteries" (mystères), into their secrets none but those duly initiated could penetrate Modern Industry rent the veil that concealed from men their own social process of production, and that turned the various spontaneously divided branches of production into so many riddles, not only to outsiders, but even to the initiated The principle which it pursued, of resolving each process into its constituent movements, without regard to their possible execution by the hand of man, created the new modern science of technology The varied, apparently unconnected, and petrified forms of the industrial processes now resolved themselves into so many conscious and systematic applications of natural science to the attainment of given useful effects Technology also discovered the few main fundamental forms of motion, which, despite the diversity of the instruments used, are necessarily taken by every productive action of the human body, just as the science of mechanics sees in the most complicated machinery nothing but the continual repetition of the simple mechanical powers
>
> Modern Industry never looks upon and treats the existing form of a process as final The technical basis of that industry is therefore revolutionary, while all earlier modes of production were essentially conservative[20]

In its most advanced form, therefore, "modern industry makes science a productive force distinct from labour and presses it into the service of capital" (Marx 1906, p 397)

Before capitalism could reach this stage of self-sustaining technological

[19] Marx 1906, p 504 The manufacturing stage needs to be seen as an essential step in the introduction of science into the productive process The application of science required that productive activity be broken down into a series of separately analyzable steps The manufacturing system, even though it continued to rely upon human skills, accomplished precisely this when it replaced the handicraftsman with a number of detail laborers In this important sense it "set the stage" for the advent of modern industry

[20] Ibid, p 532 Marx (1959) examines the vast possibilities for capital-saving innovations and improvements in an advanced capitalist economy in *Capital*, vol. 3, chaps 4 and 5.

dynamism, however, another critical condition needed to be fulfilled. Machinery cannot fully liberate the economy from the output ceiling imposed by dependence upon human skills and capacities so long as these things continue to be essential in the production of the machines themselves. In the early stages of modern industry, machines were, inevitably, produced by direct reliance upon human skills and capacities The manufacturing system responded to the demand for the new inventions by creating new worker specializations.[21] While this sufficed in the early stages of the development of modern industry, improvements in machine design and performance and increasing size eventually came up increasingly against the limitations of the human machine maker

> Modern Industry was crippled in its complete development, so long as its characteristic instrument of production, the machine, owed its existence to personal strength and personal skill, and depended on the muscular development, the keenness of sight, and the cunning of hand, with which the detail workmen in manufactures and the manual labourers in handicrafts, wielded their dwarfish implements. Thus, apart from the dearness of the machines made in this way, a circumstance that is ever present to the mind of the capitalist, the expansion of industries carried on by means of machinery, and the invasion by machinery of fresh branches of production, were dependent on the growth of a class of workmen, who, owing to the almost artistic nature of their employment, could increase their numbers only gradually, and not by leaps and bounds But besides this, at a certain stage of its development, Modern Industry became technologically incompatible with the basis furnished for it by handicraft and Manufacture The increasing size of the prime movers, of the transmitting mechanism, and of the machines proper, the greater complication, multiformity and regularity of the details of these machines, as they more and more departed from the model of those originally made by manual labour, and acquired a form, untrammelled except by the conditions under which they worked, the perfecting of the automatic system, and the use, every day more unavoidable, of a more refractory material, such as iron instead of wood—the solution of all these problems, which sprang up by the force of circumstances, everywhere met

[21] "As inventions increased in number, and the demand for the newly discovered machines grew larger, the machine-making industry split up, more and more, into numerous independent branches, and division of labour in these manufactures was more and more developed. Here, then, we see in Manufacture the immediate technical foundation of Modern Industry Manufacture produced the machinery, by means of which Modern Industry abolished the handicraft and manufacturing systems in those spheres of production that it first seized upon" (Marx 1906, p. 417)

> with a stumbling-block in the personal restrictions which even the collective labourer of Manufacture could not break through, except to a limited extent Such machines as the modern hydraulic press, the modern powerloom, and the modern carding engine, could never have been furnished by Manufacture.[22]

The vital step, therefore, is the establishment of the technological conditions which would make it possible to use machinery in the construction of machines, thus bypassing the central constraint of the old manufacturing system "Modern Industry had therefore itself to take in hand the machine, its characteristic instrument of production, and to construct machines by machines It was not till it did this, that it built up for itself a fitting technical foundation, and stood on its own feet Machinery, simultaneously with the increasing use of it, in the first decades of this century, appropriated, by degrees, the fabrication of machines proper "[23] Marx singles out, not only the new power sources which offered gigantic quantities of energy subject to careful human regulation, but also that indispensable addition to the equipment at the disposal of the machine maker, the slide rest This simple but ingenious device of Henry Maudsley replaces, as Marx perceptively notes, not any particular tool, "but the hand itself" (Marx 1906, p 408) In this sense it is a strategic technological breakthrough, fully comparable in importance to the steam engine

The improvements in the machinery-producing sector constitute a quantum leap in the technological arsenal at man's disposal They make it possible to escape the physical limitations of a tool-using culture They do this, ironically as Marx points out, by providing machines which reproduce the actions of a hand-operated tool, but do so on a "cyclopean scale "[24]

---

[22] Ibid , pp 417–18 Marx saw the improvements in the means of communication and transportation as particularly significant in pushing the productive process beyond the limitations inherent in the manufacturing system "The means of communication and transport became gradually adapted to the modes of production of mechanical industry, by the creation of a system of river steamers, railways, ocean steamers, and telegraphs But the huge masses of iron that had now to be forged, to be welded, to be cut, to be bored, and to be shaped, demanded, on their part, cyclopean machines, for the construction of which the methods of the manufacturing period were utterly inadequate" (pp 419–20)

[23] Ibid , p 420 Marx saw this process as culminating during his own time "It is only during the last 15 years (i e , since about 1850), that a constantly increasing portion of these machine tools have been made in England by machinery, and that not by the same manufacturers who make the machines" (p 408)

[24] "If we now fix our attention on that portion of the machinery employed in the construction of machines, which constitutes the operating tool, we find the manual implements reappearing, but on a cyclopean scale The operating part of the boring machine is an immense drill driven by a steam-engine; without this machine, on the other hand, the cylinders of large steam-engines and of hydraulic presses could not be made. The mechanical lathe is only a cyclopean reproduction of the ordinary footlathe, the planing machine, an iron carpenter, that works on iron with the same tools that the human carpenter employs on wood; the instrument that, on the London wharves, cuts

IV

Thus, I would interpret the Marxian position to be that it is the changing requirements of industry and the altering perception of economic needs which provide the stimulus to the *pursuit* of specific forms of scientific knowledge. But I would also conclude that the Marxian position cannot be adequately described as a demand-induced approach without doing a severe injustice to the subtlety of Marx's historical analysis.[25] For the ability to apply science to the productive sphere turns upon industry's changing *capacity* to utilize such knowledge, a capacity which Marx explicitly recognizes has been subjected to great changes over the course of recent history. Indeed, Marx himself, as I have tried to establish, devoted considerable effort to the elucidation of the factors which have shaped society's altering capacity to absorb the fruits of scientific knowledge.[26]

Nor did Marx argue that the historical sequence in which scientific disciplines actually developed was also directly determined by economic needs. For example, in discussing the relative pace of development in industry and agriculture, he states that productivity growth in agriculture had, historically, to await the development of certain scientific disciplines, and therefore came later, whereas industry progressed more rapidly than agriculture at least in large part because the scientific knowledge upon

---

the veneers, is a gigantic razor, the tool of the shearing machine, which shears iron as easily as a tailor's scissors cut cloth, is a monster pair of scissors, and the steam hammer works with an ordinary hammer head, but of such a weight that not Thor himself could wield it. These steam hammers are an invention of Nasmyth, and there is one that weighs over 6 tons and strikes with a vertical fall of 7 feet, on an anvil weighing 36 tons. It is mere child's play for it to crush a block of granite into powder, yet it is not less capable of driving, with a succession of light taps, a nail into a piece of soft wood" (ibid., p. 421; see also pp. 492–93).

[25] At one point Marx presents what one might be tempted to call a Toynbeean "challenge-response" mechanism to account for the emergence of high productivity societies. It is not true, he says, "that the most fruitful soil is the most fitted for the growth of the capitalist mode of production. This mode is based on the dominion of man over nature. Where nature is too lavish, she 'keeps him in hand, like a child in leading-strings.' She does not impose upon him any necessity to develop himself. It is not the tropics with their luxuriant vegetation, but the temperate zone, that is the mother country of capital. It is not the mere fertility of the soil, but the differentiation of the soil, the variety of its natural products, the changes of the seasons, which form the physical basis for the social division of labour, and which, by changes in the natural surroundings, spur man on to the multiplication of his wants, his capabilities, his means and modes of labor. It is the necessity of bringing a natural force under the control of society, of economising, of appropriating or subduing it on a large scale by the work of man's hand, that first plays the decisive part in the history of industry" (ibid., pp. 563–64).

[26] In this light, there is no necessary conflict between Marx's materialist conception of history and his treatment of science as a productive force under advanced capitalism. I therefore disagree with the following statement of Bober: "Marx intends to offer a materialistic conception of history. Yet he frequently stresses the power of science as a component of modern technique and production. The incorporation of science in the foundation of his theory is no more defensible than the inclusion of all other nonmaterial phenomena" (Bober 1965, p. 21).

which industry relied had developed earlier. "Mechanics, the really scientific basis of large-scale industry, had reached a certain degree of perfection during the eighteenth century. The development of chemistry, geology and physiology, the sciences that *directly* form the specific basis of agriculture rather than of industry, does not take place till the nineteenth century and especially the later decades "[27]

This strongly suggests at least some degree of independence and autonomy on the part of science in shaping the sequence of industrial change, in spite of the fact that, as we saw earlier, Marx and Engels usually emphasize the cause-effect relationships which run from industry to science If the growth in agricultural productivity is dependent upon progress in specific subdisciplines of science, and if the existence of profitable commercial opportunities in agriculture cannot "induce" the production of the requisite knowledge, then factors internal to the realm of science must be conceded to play a role independent of economic needs

Moreover, it is especially curious to find that Engels is content to state, as quoted earlier, that "from the very beginning the origin and development of the sciences has been determined by production" (Engels 1954, p 247) For Engels himself, in the *Dialectics of Nature*, had also presented a classification scheme for the sciences which emphasized a hierarchy of increasing complexity based upon the forms of motion of the matter being analyzed Increasing complexity is identified with the movement from the inorganic to the organic, from mechanics to physics to chemistry to biology [28] Engels even goes so far as to speak of an

---

[27] Marx 1968, pt 2, p 110 In *The German Ideology* Marx and Engels stated that "the science of mechanics perfected by Newton was altogether the most popular science in France and England in the eighteenth century" (Marx and Engels 1947, p 56)

[28] "Hegel's division (the original one) into mechanics, chemics, and organics, fully adequate for the time Mechanics the movement of masses Chemics molecular (for physics is also included in this and, indeed, both—physics as well as chemistry—belong to the same order) motion and atomic motion Organics the motion of bodies in which the two are inseparable For the organism is certainly *the higher unity which within itself unites mechanics, physics, and chemistry into a whole* where the trinity can no longer be separated In the organism, mechanical motion is effected directly by physical and chemical change, in the form of nutrition, respiration, secretion, etc , just as much as pure muscular movement" (Engels 1954, pp 331–32, emphasis Engels's) For Engels's entire treatment of the subject, see ibid , pp 322–408 In his book, *Herr Eugen Duhring's Revolution in Science*, Engels draws a sharp distinction between the sciences concerned with inanimate nature and those concerned with living organisms The former group of sciences (mathematics, astronomy, mechanics, physics, chemistry) are susceptible to mathematical treatment "to a greater or less degree " No such precision is possible in the sciences concerned with living organisms "In this field there is such a multitude of reciprocal relations and causalities that not only does the solution of each question give rise to a host of other questions, but each separate problem can usually only be solved piecemeal, through a series of investigations which often requires centuries to complete, and even then the need for a systematic presentation of the interrelations makes it necessary again and again to surround the final and ultimate truths with a luxuriant growth of hypotheses" (Engels 1939, pp 97–99)

"inherent sequence,"[29] which he clearly believes has structured the *historical* sequence in which nature's secrets have been progressively uncovered. But, if one accepts this intuitively plausible view, then surely there is much more to "the origin and development of the sciences" than can be accounted for by the specific demands being generated in the productive sphere Surely the historical fact that the biological sciences came to the assistance of agriculture long after the mechanical sciences were being utilized by industry is a sequence originating, not in economic needs, but in the differing degrees of complexity of these scientific disciplines Engels's formulations particularly seem to overemphasize the importance of demand-induced incentives to the neglect of supply side considerations, even though he is obviously sensitive to these supply variables in other contexts

In Engels's defense one must recall, of course, the unfinished, indeed often merely fragmentary condition of his *Dialectics of Nature*[30] It is entirely possible that, had he the opportunity, he would have resolved these apparent inconsistencies But it is expecting far too much to look to either Marx or Engels for the resolution of these deep and thorny problems We are still, today, a long way from being able to incorporate the history of science in an orderly manner into our understanding of the economic development of the Western world[31]

Conclusion

There are several possible meanings which can be attached to the statement that "the origin and development of the sciences has been determined by production "

1 Science depends upon industry for financial support

2 The expectation of high financial returns is what motivates individuals (and society) to pursue a particular scientific problem

3 The needs of industry serve as a powerful agent in calling attention to certain problems (Pasteur's studies of fermentation and silkworm epidemics)

---

[29] "Classification of the sciences, each of which analyzes a single form of motion, or a series of forms of motion that belong together and pass into one another, is therefore the classification, the arrangement, of these forms of motion themselves according to their inherent sequence, and herein lies its importance" (Engels 1954, p 330, see also Zvorikine 1963, pp 59–74)

[30] See Engels 1954, "Preface "

[31] The most ambitious attempt to fill this void is the fascinating but seriously flawed four-volume work by the late J D Bernal, *Science in History* (1971) His *Science and Industry in the Nineteenth Century* (London, 1953) is more restricted in scope and far more consistently persuasive Nevertheless, *Science in History* displays an immense erudition, and all but the most remarkably well-informed readers will learn much from, and be greatly stimulated by, its contents

4 The normal pursuit of productive activities throws up physical evidence of great importance to certain disciplines (metallurgy and chemistry, canal building and geology) As a result, industrial activities have, as a byproduct of their operation, provided the flow of raw observations upon which sciences have built and generalized

5 The history of individual sciences, including an account of their varying rates of progress at different periods in history, can be adequately provided by an understanding of the changing economic needs of society

I believe that Marx and Engels subscribed to propositions 1–4 without qualification I believe they often *sounded* as if they subscribed to the fifth proposition However, I think the preceding discussion has established that they subscribed to the fifth proposition only subject to certain qualifications—qualifications which strike me as being, collectively, more interesting than the original proposition

References

Bernal, J D *Science and Industry in the Nineteenth Century* London Routledge & Kegan Paul, 1953

——— *Science in History* 4 vols Cambridge, Mass M I T Press, 1971

Bober, M M *Karl Marx's Interpretation of History* New York · Norton, 1965

Engels, Frederick *Socialism Utopian and Scientific* Chicago Kerr, 1910

——— *Herr Eugen Duhring's Revolution in Science* New York International, 1939

——— *The Dialectics of Nature* Moscow Foreign Languages, 1954

Marx, Karl *A Contribution to the Critique of Political Economy* Chicago Kerr, 1904.

——— *Capital* Vol 1 Chicago Kerr, 1906

———. *Capital* Vol 3 Moscow Foreign Languages, 1959

——— *Theories of Surplus Value* 3 pts Moscow Progress, 1963

——— *The Poverty of Philosophy* Moscow Foreign Languages, n d

Marx, Karl, and Engels, Frederick *The German Ideology* New York International, 1947

——— *Selected Works* 2 vols Moscow Foreign Languages, 1951

Sweezy, Paul "Karl Marx and the Industrial Revolution " *Events, Ideology and Economic Theory,* edited by Robert Eagly Detroit Wayne State Univ Press, 1968

Zvorikine, A "Technology and the Laws of its Development " *The Technological Order,* edited by Carl F Stover Detroit Wayne State Univ Press, 1963

Advertising as Information

Phillip Nelson

State University of New York at Binghamton and the University of Chicago

> This haughty youth, He speaks the truth
> Whenever he finds it pays,—
> And in this case, it all took place
> Exactly as he says!
> Exactly, exactly, exactly, exactly as he says!
>
> [GILBERT AND SULLIVAN, *The Mikado*]

This paper tries to show how the major features of the behavior of advertising can be explained by advertising's information function For search qualities advertising provides direct information about the characteristics of a brand For experience qualities the most important information conveyed by advertising is simply that the brand advertises This contrast in advertising by these qualities leads to significant differences in its behavior

How does advertising provide information to the consumer? The producer in his advertising is not interested directly in providing information for consumers He is interested in selling more of his product Subject to a few constraints, the advertising message says anything the seller of a brand wishes A mechanism is required to make the selling job of advertising generate information to the consumer

The first section of this paper shows how such a mechanism exists and how its operation varies by types of market The middle three sections of the paper try to put the ideas of the first section to the test The first of these empirical sections uses the model to predict the way in which adver-

---

Helpful criticisms of an earlier draft of this article were received from Jerome Komisar, Lester Telser, Robert Van Handel, and the participants of a seminar at the University of Rochester and the Conference on the Economics of Information at the University of Chicago Financial assistance was received from the Research Foundation of the State University of New York

tising/sales ratios will differ by industry. The second section predicts how media choice will vary by industry, the third explores the impact of alternative sources of information on advertising volume. The last section of this paper examines the consequences of relaxing two assumptions made in the rest of my analysis: that consumers use optimal decision rules and that there are no legal constraints on advertising.

Information is generated by advertising because of consumer power in the product market. The nature of that power will vary significantly by the nature of the product being advertised. In "Information and Consumer Behavior" (Nelson 1970), I make a fundamental distinction between qualities of a brand that the consumer can determine by inspection prior to purchase of the brand—"search qualities"—and qualities that are not determined prior to purchase—"experience qualities." An example of a search quality is the style of a dress, an example of an experience quality is the taste of a brand of canned tuna fish. The impact of the market power of consumers on advertising will operate quite differently for search and experience qualities.

If the advertised properties of the product differ from the actual properties, the consumer will know about that difference prior to purchase in the case of search qualities. This reduces considerably—but not entirely—incentives for misleading advertising. Misleading advertising will still exist. First, consumers pay transportation costs before they spot any disparity between actual and advertised search qualities. By exaggerating the utility of the actual search good by any amount up to the transportation costs of consumers, advertising can result in greater sales to the firm than a less exaggerated advertisement. Second, there can be some search quality of a brand which cannot be successfully conveyed by advertising, for example, the "charm" of an apartment. Under these circumstances it might pay the advertiser to exaggerate some characteristic that can be specified by the advertisement. But there are costs to the advertiser of misleading advertising: he suffers a decline in his credibility for future advertisements and pays the costs of processing nonbuying customers. Because of these costs and the relative unimportance of the sources of misleading advertising, consumers can have some confidence that the advertising of search qualities bears a close relation to the truth. The advertising of search qualities provides information to the consumer, even though he attaches a probability less than one to the truthfulness of these advertisements.

In the case of experience qualities the consumer's power over advertising is much less potent than his power over search qualities. The major control that consumers have over the market for experience qualities is whether they repeat the purchase of a brand or not (Nelson 1970). This power is sufficient to authenticate any statement that has either of the following characteristics. (1) consumers' belief in the truth of that statement does not increase the profit from initial sales (as opposed to repeat sales either directly or indirectly through the recommendations of relatives and friends) or (2) it is possible for the producer at no cost to himself to make the advertised statement true rather than false If either (1) or (2) holds, the advertiser maximizes his profits by telling the truth if the consumer believes he is doing so.[1]

One of these two requirements will be satisfied by virtually all statements that specify the function of a brand, for example, "Pepto-Bismol is a remedy for upset stomachs " This statement turns customers with athlete's foot away as well as attracting customers with bellyaches Of course, it is conceivable that there are more of the latter (relative to the available number of brands) than the former, so that there is a net gain of initial sales from this statement But this is where condition (2) comes into play If this were the case, it would pay the manufacturer of Pepto-Bismal to produce at the same costs something that resembled a stomach remedy rather than something that could be used for athlete's foot It is conceivable that at that cost the firm could only produce a second-rate stomach remedy and could produce a first-rate foot balm But if everybody is buying Pepto-Bismol for upset stomachs rather than for athlete's foot, any virtues of Pepto-Bismol for athlete's foot would be irrelevant in determining Pepto-Bismol sales

Even when the advertisement of an experience quality correctly relates brand to function—by far the typical case—the consumer has quite incomplete information He would like to be able to rank stomach remedies by their utility to him Advertising provides no direct information that will help him do that job (By direct information I mean information contained in the advertising statement) After Pepto-Bismol has been correctly identified as a stomach remedy, the statement that Pepto-

---

[1] This proposition can be seen most clearly in the case of constant average production costs (AC) In that case, profit can be written

$$Pf = (P - AC)I + (P - AC)R - A, \tag{10}$$

where Pf = profits, P = price, I = initial sales, R = repeat sales, A = advertising costs, and $P(R)$ = probability of a repeat purchase given an initial purchase Advertising costs are assumed constant, since they should be invariant with respect to the truth or falsity of the advertising message Then false advertising will not increase Pf

$$Pf = (P - AC)I + (P - AC)I\,P(R) - A \tag{11}$$

If $(P - AC)I$ is constant, profits increase with the probability of a repeat purchase, and this probability increases with the truthfulness of an advertisement $(P - AC)I$ will be a constant over the relevant decision variable whenever either condition (1) or (2) holds If condition (1) holds, $(P - AC)I$ is a constant varying the truthfulness of the statement but holding the character of the product constant If condition (2) holds, $(P - AC)I$ is a constant holding the statement constant but varying the character of the product

Bismol is most soothing is without information content. Its producers have an incentive to say so even if it were the least soothing of stomach remedies.

The miniscule amount of direct information from advertising for experience qualities gives the consumer an incentive to extract any conceivable indirect information that would help. Such indirect information is available from advertising. The consumer can learn that the brand advertises. I contend that this is the useful information that the consumer absorbs from the endorsements of announcers, actors, and others who are paid for their encomiums. These and other advertisements for experience goods have no informational content. Their total informational role—beyond the relation of brand to function—is simply contained in their existence. The consumer believes that the more a brand advertises, the more likely it is to be a better buy. In consequence, the more advertisements of a brand the consumer encounters, the more likely he is to try the brand.

Why are advertised brands better buys? The answer to that question is not self-evident. If advertising were distributed at random among brands for a product, it would indeed be newsworthy that a brand advertised, but that news would decrease the probability of a consumer's trying the advertised brand. Advertised brands would cost more by the cost of the advertising and hence would be worse buys.

It is my contention, however, that advertising is not distributed at random among brands of a product. Heavily advertised brands are likely to provide a lower P^* (price per unit of utility of the brand)[2] to the average consumer than less heavily advertised brands of the same product. (To give unambiguous meaning to that measure, I assume that consumers all have the same income and utility functions.)

First of all, firms vary in their efficiency in producing the utility that consumers seek. Some firms produce brands that yield more utility to the consumer for a dollar of production cost than do other brands. In general a firm that has lower costs relative to the utility of its brand than other firms will find that it pays to expand its output by both increasing advertising expenditures and decreasing P^*. This behavior of firms by efficiency generates a negative association between advertising and P^* by brands.

Two crucial propositions are contained in the previous paragraph. The proposition—that to sell more the firm will offer a lower P^*—obviously holds for the usual demand curve, which is negatively sloped. More critically, this proposition also holds for the relationship between demand and price when advertising expenditures adjust optimally to price. Even though this demand curve can have a positively sloped portion (Demsetz 1959), Schmalensee's (1972) analysis implies that even this demand curve will be declining where it counts—in the neighborhood of equilibrium.[3] The proposition that there is a reverse association between quantity produced and P^* can be put to the test. If this proposition were true, it would pay a brand to advertise its rank in its product class more, the higher that rank. Do leading firms advertise their significance more than nonleading firms advertise their lack of leadership? To answer that question, I looked at three issues of *Life* magazine for March 1972 (*Life* 1972). I found six brands claiming that they were at largest brands for their respective product classes. I found none advertising a lower rank. There are far fewer brands that are first in their product class than are not first (even confining ourselves to the sample of brands that advertise in three issues of *Life*). If brands advertised their rank in their product class independently of what that rank happened to be, the probability of an advertisement which advertised rank advertising that the rank is first would be far less than .5. In consequence the probability of our observed results occurring by chance is less than one sixty-fourth.[4]

---

[2] This definition of P^* is an attempt to obtain a utility adjusted price. For simple additive utility functions, P^* does the job. For other utility functions, different definitions of P^* might be required. For purposes of my analysis this variation in the definition of P^* is not important as long as a utility adjustment of price is possible.

The second proposition—that the firm that wishes to sell more will advertise more—also has a strong a priori defense. Advertising and $-\Delta P^*$ are the two inputs yielding higher quantity demanded. From production theory one anticipates that if the "price" of the two inputs were constant, an increase in the output is usually most profitably achieved by an increase in both inputs. But in this case the "price" of the inputs does not remain constant as one increases quantity. Quantity is the "price" of $-\Delta P^*$, that is, the loss to the firm in lowering price to increase quantity demanded is proportional to that quantity. In consequence, one would expect advertising expenditures to increase because both output increases and the price of the other factor of production increases The evidence is overwhelming that advertising increases as sales increase.

---

[3] Using Demsetz's (1959) terminology, the average revenue curve allowing advertising to adjust optimally to a change in price is the *mutatis mutandis* average revenue curve (MAR). For MAR to be rising or constant $MMR \geq MAR$, where MMR = marginal *mutatis mutandis* revenue curve. For firms to have positive or zero profits $MAR \geq AC$, where AC = average production and advertising costs and MC = marginal production and advertising costs, but $MMR = MC$. Hence, $MC \geq AC$, which implies that AC is rising or constant. But Schmallensee (1972) has shown that in equilibrium AC must be decreasing. Hence, MAR must be decreasing in equilibrium.

[4] This test would be biased if brands that were not number one advertised themselves as major producers more than those brands which are number one. But this bias is produced by the very proposition I am testing, a proposition that would explain why number one brands choose more explicit rank advertising than others. A more serious problem with this approach is that the data do not support another implication of rational consumer behavior: that firms that advertise the most would advertise that fact. I think the primary reason that the quantity of advertisement is not advertised is that advertising has been so frequently attacked that it is in disrepute. These attacks make it difficult for consumers to understand the ultimate rationalization of their own behavior. (This problem is discussed in more detail in the "Deceptive Advertising" section of this paper.) In consequence, brands do not advertise their advertising rank, though they do advertise that they have advertised, e.g., "As Advertised in *Life*."

There are other reasons for a consumer to respond positively to advertising What is a high-utility brand for some consumers will be a low-utility brand for others Advertising has the seemingly magical property that those whose tastes are best served by a given brand are those most likely to see an advertisement for that brand Advertisers choose the media in which they advertise in part to maximize the repeat-purchase probability for their brand In consequence the producer distributes his advertisements among media so that his message is seen by those who are most likely to repeat purchase his brand An esoteric, high-price soup gets advertised in the *New Yorker*, while Campbell's soup displays its wares in *Good Housekeeping*

A third factor operates to increase the reputability of advertised brands The full analysis of this factor is consigned to the Appendix, but the essence of the argument is simple enough Advertising increases the probability of a consumer's remembering the name of a brand Those brands with the highest probability of repeat purchase have the greatest payoff to improved consumer memory In consequence, brands which provide the highest utility have the greatest incentive to advertise

A feature of the previous discussion must be emphasized Nowhere in that analysis was it necessary to assume that consumers respond positively to advertising because the brands that advertised the more heavily were better buys The self-interest of consumers to respond to advertising only if it increased their utility produces an additional guarantee that highly advertised products will provide higher utility to the consumer Suppose there were some products for which firm self-interest dictated a negative relationship between advertising and utility per dollar Consumer self-interest would make the consumers respond negatively to advertising for these products In consequence, firms would not advertise such products

In this analysis of the indirect information contained in advertising, the focus has been on experience qualities The same analysis can be used to show that advertising increases the reputability of brands which are dominantly search goods However, reputability will play a much smaller role in the advertising for search qualities, because the consumer can obtain so much direct information about these qualities from advertisements and direct inspection The possession of this direct information reduces the payoff to both consumers and advertisers of advertisements' increasing the reputability of a brand In my subsequent analysis I shall assume that this difference between the character of advertising for search and experience qualities is so large that the advertising for experience qualities is dominantly indirect information and the advertising for search qualities is dominantly direct information

Marginal Revenue of Advertising to the Producer

To maximize profits a producer will advertise to the point where his marginal revenue of advertising is equal to his marginal cost of advertising One expects differences in the behavior of this marginal revenue for search and experience qualities This difference in turn will produce a differential effect on the quantity of advertising purchased by producers of search and experience qualities

To demonstrate this effect one needs a theory to generate the marginal revenue of advertising For search qualities, advertising will increase sales when and only when it gives the consumer information that he did not have before I will assume that all the information that advertising conveys about a search brand is contained in any given message The revenue (R) generated by advertising will, therefore, equal·

$$R = QPNG_1, \qquad (1)$$

where N = the number of potential customers, that is, those who would buy the brand if confronted with the appropriate signal (in this case, one or more advertising messages), Q = average quantity purchased per customer, P = price, and G_1 = the proportion of potential customers with one or more advertising messages Of the four components of R only G_1 varies by advertising intensity The behavior of G_1 has been examined by Stigler (1961) and Gould (1970), but there is a problem with their analysis that makes reexamination necessary

As Stigler recognizes, there are three relevant processes at work Consumers acquire advertising information, they leave and enter the market, that is, they are mobile, they forget advertising information Stigler asserts that the latter two processes produce identical results in the advertising market, except for possible differences in their intensities I think he is wrong In the mobility case, the most reasonable simple assumption is that the probability of leaving the market is independent of the number of advertising messages a consumer knows In the forgetting case, however, the simplest reasonable assumption is that the number of messages forgotten about a brand is directly proportional to the number of messages known about a brand

I expect, then, mobility and forgetting to be quite different probability processes Optimally, one should combine the two processes in the same model Such a model is easy to construct, but I have not been able to solve the equations that the model generates Instead I have developed two different models In one there is mobility, but no forgetting, in the others there is forgetting, but no mobility Because it is simpler and probably more important, I concentrate on the forgetting model This is simply a matter of economizing on space My results do not change when I use the mobility model [5]

---

[5] My analysis of the mobility model is available upon request

In my forgetting model I assume that the rate at which consumers confront an advertising message (c) is a constant for all consumers. I also assume that the forgetting rate is directly proportional to the number of advertising messages a person knows. The differential matrix of this stochastic process is

$$F = \begin{bmatrix} -c & c & & & \\ a & -(c + a) & c & & \\ & 2a & -(c + 2a) & c & \\ & & 3a & -(c + 3a) & c \end{bmatrix}, \qquad (2)$$

where the state space is $(0, 1, \ldots, n)$ advertising messages about a brand and a is the forgetting rate when one knows one advertising message.

I shall try to abstract out of this stochastic process the relationship betw[illegible] the present value of revenue and advertising intensity. This present value of revenue will change over time as the process unfolds. Of these present values both the simplest and the most useful is the one calculated after the stochastic process has settled down to a steady state. Then the present value will merely be the present value of a constant stream of returns. It is this steady-state condition that I will examine.

As is well known (e.g., Karlin 1966), the steady-state distribution (λ) can be obtained by solving

$$\lambda F = 0. \qquad (3)$$

In words, the net movement into any steady state must be zero or it is not a steady state. The solution

$$\lambda_i = \frac{d^i}{i!} e^{-d} \qquad i = (0, 1, \ldots, n), \qquad (4)$$

where $d = c/a$. In particular the probability of having one or more advertising messages is

$$1 - \lambda_0 = 1 - e^{-d}. \qquad (5)$$

Equation (5) is G_1 in equation (1) and hence is relevant in determining the revenue from advertising in the case of search.

However, G_1 will not be relevant in determining the revenue from advertising for experience qualities. I have assumed that the information the consumer obtains from advertising for experience qualities is that the brand is advertised and is, hence, more likely to be a better buy. In the face of experience qualities, a consumer's optimal strategy is to try a fixed number of brands, say r, and then continue to use the best of the set (Nelson 1970). A consumer to whom advertising for experience qualities provides favorable information will limit his sampling to those brands for which he has received the most advertising messages. Say he has currently received m messages for the rth most advertised brand in his own sample of messages. Assume further that in case of a tie in the number of advertising messages the consumer receives, the consumer prefers the brand whose advertising messages he has most recently received [6] Then a consumer will try a brand if he receives m messages about the brand, and any messages beyond m are redundant

For experience qualities one needs a more general form of equation (1) The revenue generated by advertising is

$$R = QPNG_m \tag{6}$$

where G_m = the proportion of potential customers with m or more advertising messages, where m is the critical number of advertising messages required for a consumer to sample a brand (To put the revenue in present-value terms, all one need do is multiply the right side of eq [6] by $1/w$—where w equals the discount rate) For the forgetting model, G_m is

$$1 - \sum_{i=0}^{m-1} \lambda_i = 1 - \sum_{i=0}^{m-1} \frac{d^i}{i!} e^{-d} \tag{7}$$

The amount of advertising can be measured by the rate at which customers acquire advertising information—or c The marginal revenue to the producer of advertising is, therefore

$$MR_A = QPN\frac{\partial G_m}{\partial c} \tag{8}$$

Assume that QPN are the same for advertising for search and experience qualities [7] The MR_A will differ between search and experience qualities because $\partial G_m/\partial c$ differs by m (I will assume, for simplicity, that all consumers are faced with the same m in the advertising for any given brand, though this assumption is contrary to fact)

$$\frac{\partial G_m}{\partial c} = \frac{d^{m-1}}{a(m-1)!} e^{-d} \tag{9}$$

---

[6] A more realistic assumption would be that in a case of a tie at m messages, the probability of the consumer's sampling the brand is the inverse of the number of tied brands But this more realistic assumption increases considerably the complexity of the presentation without adding any new insight into the process

[7] This assumption is contrary to fact, but its conflict with reality should not produce serious problems in this particular case In Nelson (1970) I showed that there should be more monopoly power in the market for experience goods than in the market for search goods Assume the same total market size—and there is no reason to assume that this varies systematically between search and experience goods Then average firm size should be larger for experience goods than for search goods and there should be fewer firms If marginal production and advertising costs were constant with respect to size, these two differences should cancel one another out, i e , total advertising expenditures for the industry would not vary through variation in the size of firms within the industry

For search qualities $m = 1$, for experience qualities $m \geq 1$ It can be easily shown that

$$\frac{\partial G_m}{\partial c} > \frac{\partial G_1}{\partial c},$$

where $m > 1$ and $G''_m < 0$[8] In other words, over the declining portion of $\partial G_m/\partial c$, $\partial G_m/\partial c$ is always greater for experience qualities than for search qualities (or equal when $m = 1$) Since only the declining portion of the marginal revenue curve is relevant for firm decision making when marginal costs are constant or increasing, firms advertising experience qualities will advertise more than firms advertising search qualities when they both advertise

Can the marginal cost of production and advertising always be above the marginal revenue curve of advertising for experience qualities? This is possible for any given brand of a product (which will then not advertise), it is impossible for the market in general If for all brands, the marginal cost curve is higher than the marginal revenue curve for an m as low as two, then the market will make the minimum m one The marginal revenue curve for search and experience will then be the same only in this polar case Of course the industry's total advertising expenditure are independent of the number of firms as long as marginal costs of production and advertising are constant

How does one test the proposition that there will be more advertising of experience than of search goods? In "Information and Consumer Behavior" (Nelson 1970), it is shown that goods could be classified successfully by whether the quality variation was ascertained predominantly by search or by experience, and the respective goods were called "search goods" and "experience goods" To take account of some ambiguity in the classification procedure in Nelson (1970), two alternative classifications were used They are again used here, together with a third classification (Classification II with cameras moved from the search to the experience category) This third classification is motivated by the discussion in Nelson (1970, p 320 n) The commodities in these classifications are presented in table 1 The following mean advertising ratios were

---

[8] $$G''_m = \frac{1}{a^2} d^{m-2} \frac{1}{(m-2)!} e^{-d} \left(1 - \frac{d}{m-1}\right)$$

Since $d > 0, m \geq 2$, then $G''_m < 0$ implies $\dfrac{d}{m-1} > 1$, so

$$G'_m - G'_1 = \frac{1}{a} e^{-d} \left[\frac{d^{m-1}}{(m-1)!} - 1\right]$$

But $\dfrac{d}{m-i} > \dfrac{d}{m-1}$ for $m - 1 > i > 1$. Hence, $\dfrac{d}{m-1} > 1$ implies that $\dfrac{d^{m-1}}{(m-1)!} > 1$

Hence, $G'_m > G'$ when $G''_m < 0$.

TABLE 1

ADVERTISING/SALES RATIOS IN CONSUMER CATEGORY BY INFORMATION CATEGORY*

Good and Information Category	Advertising/Sales Ratio, 1957	Good and Information Category	Advertising/Sales Ratio, 1957
Experience durable		Experience nondurable	
Books	2 702	Beer	6 872
Paints†	1 450	Wine	4 395
Tires	1 385	Liquor	2.408
Appliances	3 296	Dairy products	1 885
Motorcycles and bicycles	1 078	Grain mill products	1 695
Motor vehicles	0 907	Cereals	4 845
Motor vehicle parts and accessories	0 700	Bakery	2 803
Professional and scientific instruments‡	2 086	Sugar	0 280
Clocks and watches	5 629	Confectionery	3 543
Communications equipment	2 034	Miscellaneous foods	4 073
Average	2 177	Cigars	2 370
Average total experience	3 427	Other tobacco	5 429
Search goods		Drugs	10 280
Knit goods	1 075	Soaps	7 938
Carpets	2 052	Perfume	14 723
Hats	2 124	Petroleum refining	0 507
Millinery	0 326	Meats	0 610
Men's clothing	0 928	Periodicals	0 304
Women's clothing	1 263	Average	4 085
Miscellaneous apparel	1 269		
Furniture	1 451		
Footwear	1 326		
Leather goods	1 204		
Jewelry§	2 202		
Costume jewelry§	2 498		
Average	1 395		

*Goods classified by Classification III data from Telser (1964)
†Search good in Classification I
‡Search good in Classifications I and II
§Experience good in Classification I

obtained

$$\text{Classification I} \quad \bar{X}_e = 0347, \bar{X}_s = 0123, D = 0234, t = 3.67,$$

$$\text{Classification II} \quad \bar{X}_e = 0348, \bar{X}_s = 0137, D = 0137, t = 3\ 16,$$

$$\text{Classification III} \quad \bar{X}_e = 0343, \bar{X}_s = 0132, D = 0211, t = 3\ 40,$$

where $\bar{X}$ = mean advertising sales ratio, e = experience goods, s = search goods, D = difference of the means, t = test result of the hypothesis that $D = 0$. For all these classification procedures the differences are large, and hence are economically as well as statistically significant

The chief problem with this test and subsequent tests to be made is the possible ambiguity in the classification of goods into the two information categories search and experience Alternative classification procedures do not appear to have a large effect Of course, random errors in classification bias result toward zero observed differences in the means, so they do not vitiate the procedure The big problem in classification is an unconscious (or conscious!) tendency for the classifier to produce the results he is seeking Hopefully, the present classification, which was developed prior to the present tests, will be found to be valid [9]

In consequence our results support the hypothesis that producers of experience goods advertise more than producers of search goods This result is important because it in turn supports our fundamental behavioral proposition that advertising of experience qualities increases sales through increasing the reputability of the seller, while advertising of search qualities increase sales by providing the consumer with "hard" information about the seller's products

There is another important feature of these results They are precisely the opposite of what would be anticipated given an obvious alternative theory of the way in which advertising for experience qualities operates As I discussed in the first section, there is, indeed, some "hard" information conveyed by advertising for experience goods A consumer is able to relate brand to function For search goods advertisements can provide far more information than the relation of brand to function

If advertising were solely concerned with distribution "hard" information, there should be more advertising measured in dollar terms for search goods than experience goods, simply because there is more "hard" information that can be conveyed about search qualities Think, for example, of an advertisement saying simply, "Bayer is an aspirin " Then think of an advertisement saying, "Jonathan Logan is a dress," and showing a picture of a Jonathan Logan dress The latter advertisement takes up more space and, hence, is more expensive As we have seen, however, the data produce results precisely contrary to this prediction This strongly suggests that the information content of advertising for experience goods goes far beyond the relation of brand to function—another reason for believing that advertising for experience goods provides information about the reputability of brands

---

[9] Obviously, the simple test used in the text makes no attempt to control for other variables Instead I have tried to present as many simple tests as I can in the belief that the many biases to which any one test is subjected (of which excluded variables is only one example) will not operate consistently in all my tests One example of a possible bias is the close relationship between search goods and style goods (This relationship is not accidental but engendered by the search—in contrast to the experience—process) It would appear, however, that the bias induced by the association of search and style runs against the test results reported in the text The evidence we have (e g , Telser 1964) suggests a positive association between advertising in an industry and the number of new brands in the industry The same logic which generates that association would make style changes—a characteristic of style goods—generate additional advertising That search goods tend to be style goods could hardly explain the low levels of advertising in search goods

This last paragraph does not imply that the relation of brand to function is a nonexistent activity of advertising for experience goods. The implication is only that the effects of this activity are swamped by the effects of advertising to increase the reputability of experience brands Why? Simply because most advertising of experience goods occurs with such intensity that most people already know the relation of brand to function.

This explanation is itself open to test The smaller the advertising intensity of an experience good, the more likely that the advertising conveys new information to the consumer of the relation of brand to function We have just seen that when advertising for experience goods is providing that kind of information the optimal page size of an advertisement ought to be smaller than the optimal page size for search goods, since the latter advertisements have more information to convey

When, however, advertising for experience goods is designed primarily to increase the reputability of a brand, two forces will tend to increase the page size of advertisements for experience goods relative to this page size for search goods First, even with the same amount of "hard information" and even when the consumer sees both, a larger advertisement should increase sales for experience goods more than a smaller advertisement The larger advertisement enhances the reputation of the advertiser more than does the smaller advertisement (When consumers select brands by volume of advertising, they will choose in terms of the dollar volume of advertising they see rather than simply the number of advertising messages) Under these same conditions, the smaller and the larger advertisement for search goods would produce the same increase in sales

Second, the probability of a consumer seeing an advertisement varies more with page size for experience goods than search goods As we shall see in the next section, the marginal return to the consumer is small for advertisements whose sole additional information is increases in the reputability of a brand In consequence, such advertisements must depend much more on catching a disinterested consumer's eye than advertisements that consumers seek out Large page size tends to accomplish that objective (by way of contrast, classified advertisements have both large marginal returns to the consumer and exceedingly small page sizes)

One can, therefore, predict that as the volume of advertising increases by brand, the difference between the average page size of advertisements for experience and search goods increases I tested this proposition by examining the national brand advertisements in the *New Yorker* for all of 1965 (This magazine was chosen because it is one of the few that has a lot of advertisements of both search and experience goods) I measured the volume of advertising by the total pages purchased in the *New Yorker* by brand I then divided this volume of advertising into four categories. As table 2 shows, there is a perfect rank correlation between the differences in page size of experience and search goods and the volume of

TABLE 2

Average Page Size of Advertisements in the *New Yorker* in 1965 by Brand by Volume of Advertising, and by Information Category

	Volume*							
	< 2		2 < 5		5 < 10		≥ 10	
	No Brands	Average Page Size	No Brands	Average Page Size	No Brands	Average Page Size	No Brands	Average Page Size
1 Search	138	553	123	786	75	895	28	789
2 Experience minus books†	122	466	141	730	127	888	73	1 008
3 Experience	411	253	146	752	128	883	74	1 036
Differences								
(1) – (2)		087		056		007		– 219
(1) – (3)		300		034		012		– 247

*Total pages of advertisements in the *New Yorker* for a brand

†Books omitted in (2) included in (3) because the behavior of books dominates the < 2 volume category

advertising. The probability of this perfect ordering being produced by chance is one twenty-fourth In consequence these results are significant at the 5 percent level.[10]

These data yield other striking results The average page size for the smaller volume advertisers is smaller for experience goods than for search goods, and this result is reversed for large-volume advertisers In the class of brands that advertised two pages or less in the *New Yorker* in 1965, the difference in average page size between search goods and experience goods was 087 The standard deviation of that difference was 037 Hence this difference is significant at the 5 percent level [11] For big-volume advertisers (10 or more pages in the *New Yorker* in 1965), the difference in average page size between search and experience goods was – 218, the standard deviation of that difference was 059 This difference is significant at the 01 percent level

These results explain a phenomenon inexplicable with the simple information models previously used in the analysis of advertising It is traditional in advertising to have a particular brand sponsor a network television show This results in three or more commercials during that period for the same brand Often the information content of these commercials does not differ Since the audience is quite stable over the course of the program, this advertising behavior makes no sense if the goal of the firm is to maximize the number of people who have one or more bits of information about the brand Obviously a random distribution of the brand's commercial does a better job than this bunching procedure But this bunching makes sense in terms of the goal of most television advertising to increase the reputability of a brand of an experience good

Marginal Revenue of Advertising to the Consumer

In the usual analysis of advertising, the consumer plays a rather passive role Our analysis, on the contrary, visualizes consumer decision in the quest for information as a central determinant of the behavior of advertising One crucial variable governing consumer decisions is the marginal revenue to the consumer in confronting an advertisement This marginal revenue has to be greater than the time cost to the consumer in order for him to examine an advertisement This places a constraint on both the amount and type of advertisement a brand will use

Our model suggest that there should be a systematic difference between the marginal revenue to the consumer for experience goods and the marginal revenue for search goods. The marginal revenue of information to the consumer for a good is the expected improvement in his utility as a result of using this information in his sampling of brands of this good. This expected improvement can be analyzed as the product (in the mathematical sense) of two components: (1) the standard deviation of the utility distribution of brands (i.e., the distribution of P^*) for that good, (2) the expected improvement in standard deviation units of utility that can be obtained by using that information. The first factor depends on the nature of the market for the good. The second factor depends on the nature of the information available about the good. Since the focus is on information in this analysis, it is not surprising that the second factor is the agent of the expected difference between the marginal revenue of advertising to the consumer for search and experience goods, respectively.

---

[10] I ran this test both including and excluding advertisements for books The reason for this strategy was that book advertisements dominate the small-volume advertisement category for experience goods The perfect rank correlation was produced by both procedures

[11] If books are included in experience goods, the difference is 247 and its standard error 028 The results are significant at the 01 percent level

But the first factor cannot be ignored. The nature of the market is also a function of the nature of the information (Nelson 1970). Indeed, the standard deviation of the utility distribution of experience and search goods might well be systematically different. However, the nature of that difference is not at all clear. Two forces operate in opposite directions on the relative magnitude of this standard deviation of search and experience goods. On the one hand, competitive pressure reduces the standard deviation of utilities about which consumers agree more for search goods than for experience goods (Nelson 1970). On the other hand, the greater sample size that consumers use for search goods will increase the standard deviation in utility of characteristics about which consumers disagree for search goods relative to experience goods.[12]

It is impossible to say, a priori, which of the two forces analyzed will dominate in determining the relative variance of the utility of search and experience goods for a given consumer. It would be surprising, however, if this mixed effect could outweigh the impact of the difference in the information characteristics of search and experience goods. In analyzing this difference I will assume that the payoff to the same information would be the same for search and experience goods, that is, the utility distributions by brand for a given customer are the same for search and experience goods

---

[12] A brand that specializes in satisfying an extreme taste will become the most-preferred brand of those with that taste at the expense of becoming one of the least-preferred brands of those with average taste and tastes at other extremes. The smaller the sample size of brands (assuming that it is greater than one), the smaller the payoff to such specialization. Sample sizes as small as two screen out least-preferred alternatives quite effectively, but do not effectively discriminate between most-preferred brands and brands nearby on a consumer's utility scale. Any increase in sample sizes beyond two increases the payoff to being the most-preferred brands for some people considerably, and only changes the cost of being the least-preferred brand for other people mildly. Hence, greater sample sizes of brands encourage greater variance among brands in the characteristics about which consumers disagree. In a simple model of this process, a sample size of two produces a brand distribution concentrated at the median of the distribution of consumers by their most-preferred quality, an infinite sample size produces a brand distribution that corresponds exactly with that distribution of consumers. (An analysis of this model is available upon request.)

The big difference in the character of the advertisements of search and experience goods is that advertisements for experience goods are dominantly "soft" or indirect information Consumers use "soft" information for experience goods because they have no option The primary information content of advertisements for experience goods is the information that the brand advertises For search goods, on the other hand, the consumer does have the option of using "hard" or "soft" information But one of the basic assumptions of our model is that consumers respond to the direct information contained in the advertising message—"hard" information—for search goods This implies that for search goods this "hard" information has a greater value to the consumer than "soft" information I predict, therefore, that the marginal revenue of advertisements to the consumer will be greater for search goods than for experience goods

To test this proposition I look at the distribution by media of advertising for experience and search goods Consumers will be willing to look at advertisements as long as the marginal revenue to them of so doing is greater than their marginal cost The marginal cost to the consumer is dominantly a time cost This time cost will vary by the alternatives use of the time used in watching the advertisement

Many advertising media take advantage of severe restrictions on the alternative time uses of the consumer Television and radio intersperse advertisements with programs the consumer wants to hear By listening to the commercials the consumer sacrifices only those activities he can do during the short time period of the commercial The cost to the consumer of looking at billboards and transit posters are only the activities he can do while confined to a moving vehicle Only advertisements in newspapers and magazines demand of the consumer the sacrifice of the best of his possible alternative time uses For newspapers and magazines he can arrange the time saving from not looking at the advertisements any way he chooses

Newspapers and magazines have the added property that the consumer can look at the advertisements as many times as he wishes We would expect some positive relationship between the marginal revenue to the consumer of an advertisement and the number of times he will want to look at that advertisement Hence, we would expect advertisements with a high marginal revenue to the consumer to concentrate in newspapers and magazines relative to the other advertising media, we would expect just the contrary behavior for advertisements with low marginal revenue to the consumer. We, therefore, predict that there will be more advertisements for search goods in newspapers and magazines relative to the other advertising media compared to the same behavior for experience goods

I test this proposition by comparing the distribution of advertising for

TABLE 3

RATIO OF TELEVISION NETWORK TO MAGAZINE ADVERTISING BY GOODS, 1966

Goods	Ratio*
Experience	
Automobiles	1 93
Foods	2 35
Toiletries	2 46
Tobacco	2 61
Drugs	2 58
Search	
Apparel	1 59
Household furniture	0 84

*Includes only those goods tabulated for both Alcohol not included because of prohibition on network advertising of hard liquor (U S Bureau of the Census 1967)

TABLE 4

RATIO OF LOCAL TO NATIONAL ADVERTISING BY LOCAL MEDIA, 1966

Media	Ratio*
Newspapers	4 1
Radio spot	2 2
Television spot	0 5
Outdoor	0 6
Transit	1 0

*U S Bureau of the Census (1967)

magazines against television network advertising for different kinds of consumer goods The results overwhelmingly support the hypothesis Look at the ratio of advertising expenditures for television to these expenditures for magazines for 1966 as reported in table 3 Every one of the experience goods has a higher ratio than the search goods The geometric mean of this ratio for experience goods is 2 45, while for search goods the geometric mean is 1 16 This difference is statistically significant at the 1 percent level, $t = 4\ 2$

We can have more than usual confidence in this test, because we can test the test Using quite different—but highly persuasive—evidence, Ferguson (1963) has established that local advertisements have a higher marginal revenue to the consumer than national advertisements (Ferguson's result is predicted by our theory Local advertising provides for the most part information that can be checked prior to purchase price and where to find things) If our tests were indeed a test for variation in the marginal revenue to the consumer, we should find the ratio of local to national advertisements higher for newspapers than for all other local media. Look at table 4 for the ratio of local to national advertising by local media While there are too few observations to hope for statistical significance, the differences in the order of magnitude involved are huge This evidence is highly suggestive

Alternative Sources of Information

Thus far our analysis has concentrated on consumer behavior when faced with two alternatives to sample (whether by experience or search) at random or with the aid of advertising Advertising is not the only aid to sampling that the consumer can employ The most obvious source of assistance that consumers have is the recommendation of relatives and friends and consumer magazines It seems reasonable to suppose that for experience goods consumers believe these recommendations constitute better information than advertising [13] Certainly these recommendations must sometimes provide better information or they would never be used The consumer is surrounded by a sea of advertising whose information is obtained by the consumer at virtually no cost to himself The consumer, then, always has the alternative of using advertising as his guide Instead he sometimes uses the guidance of relatives and friends Hence, that information must sometimes be better information The more consumers use relatives and friends, the less they respond to advertising

I showed in "Information and Consumer Behavior" that guidance was used more for goods for which there is a low frequency of purchase than goods for which there is a high frequency of purchase both for experience and search goods (For simplicity I will henceforth call low frequency of purchase goods "durables" and high frequency of purchase goods "nondurables," though the essence of their behavioral difference is contained in the frequency of purchase) This leads to the reverse prediction for advertising greater advertising for nondurables than for durables for both experience and search [14]

However, this relationship should be far stronger for experience goods than for search goods First, the level of guidance is far less for search goods than for experience goods In consequence, guidance should have less impact on the demand for advertising for search goods Second, advertising information is better for search goods than it is for experience goods, whereas the information of relatives and friends is worse for search goods than it is for experience goods [15] Hence, advertising information will compete more effectively with the information of guides in the case of search goods

---

[13] For search goods the case is not clear at all A consumer might well regard a picture of a dress contained in an advertisement as better information than a friend telling her that the dress is pretty

[14] Durables and nondurables will differ by another important characteristic as well The value of advertising information to the consumer for a given number of advertising messages will be greater for durables than nondurables However, this consequence will produce ambiguous effects on advertising expenditures The lower value of advertising information to the consumer for nondurables will reduce the number of advertising messages that consumers receive for nondurables, but the low value of advertising information to the consumer will also increase the expenses to the producer for every message that the consumer does receive

I find a significantly higher average advertising sales ratio for nondurable than durable experience goods no matter which of the alternative classification procedures I use for experience goods

$$\text{Classification I} \quad \bar{X}_{DE} = 2\ 293, \quad \bar{X}_{NE} = 4\ 085, \quad t = 1\ 92,$$

$$\text{Classification II} \quad \bar{X}_{DE} = 2\ 187, \quad \bar{X}_{NE} = 4\ 085, \quad t = 1\ 97,$$

$$\text{Classification III} \quad \bar{X}_{DE} = 2\ 177, \quad \bar{X}_{NE} = 4\ 085, \quad t = 2\ 028,$$

where $\bar{X}_{DE}$ = arithmetic mean advertising/sales ratio for durable experience goods, $\bar{X}_{NE}$ = arithmetic mean advertising/sales ratio for nondurable experience goods, t-values are computed for tests of the difference between two means and are significant at the 5 percent level with a one-tail test procedure For search goods the sample size is too small for any significance test, but the cursory evidence points to the reverse relationship higher advertising/sales ratio for durables than for nondurables

With British data, Doyle (1968) also finds an inverse relationship between durability and advertising/sales ratios But since Doyle did not make any distinction between search and experience goods, he did not observe that that relationship was confined to experience goods

Doyle (1968) observes one more relationship that can be explained by advertising's role as information He finds a significantly negative relationship between advertising/sales ratios by product and the unit price of that product, where the unit is defined roughly as the quantity that is ordinarily purcased at one time, for example, one car, one pound of coffee, etc A low unit price will mean some combination of the following high frequency of purchase or low total expenditure on the commodity

I have already shown why there should be a close relationship between high frequency of purchase and advertising expenditures Because the classification is so broad, however, the durability measure does not catch all of the effect of frequency of purchase on advertising So even when durability is explicitly considered in the multiple regression, part of the relationship between unit price and advertising will be attributable to the relationship between price and frequency of purchase

The relationship between unit price and total expenditures on a product will also yield a negative association between unit price and advertising The greater the total expenditures of consumers for a product, the greater the anticipated variance of utility to consumers of brands of that product In consequence, when consumers spend a lot for a commodity they will tend to use better—but more expensive—information than advertising· to wit, the guidance of relatives and friends or consumer magazines Doyle's evidence, then, provides additional support for my explanation of the impact of durability on the advertising/sales ratio Of course, this is hardly conclusive evidence Only the cumulative impact of additional studies can produce evidence that could be even remotely so characterized

---

[15] It can be demonstrated that the correlation of consumer preferences is lower for search goods than for experience goods

Deceptive Advertising

In the preceding pages I have given scant attention to deceptive advertising, though, obviously, deceptive advertising exists We have seen that some deception will occur even in the case of search qualities, where the consumer has the most market power over the content of advertisements But the amount of deceptiveness in advertising can be easily exaggerated if one simply looks at the incentives of advertisers to deceive without considering the incentives of consumers not to be deceived The circumstances under which advertisers have the greatest incentives to deceive if consumers believed them are precisely the circumstances under which consumers would be least inclined to believe advertising Deception requires not only a misleading or untrue statement, but somebody ready to be misled by that statement

One possible source of deceptive advertising is consumer confusion As long as consumers followed the decision rule believe an advertisement for experience qualities when it tells about the functions of a brand, do not believe the advertisement when it tells how well a brand performs that function—the consumer will be rarely deceived But there is no guarantee, of course, that consumers will always use that decision rule

There is another important source of deceptive advertising the law Whenever a law on advertising practices is moderately enforced, deceptive advertising is sure to occur Take, for example, the law prohibiting the mislabeling of the fabric content of clothing If that law is sufficiently enforced, consumers will believe that a clothing label is usually correct This will provide an incentive for some manufacturer to mislabel—unless the law is enforced so vigorously that nobody gains from breaking it—a nonoptimal level of law enforcement (Becker 1968, Stigler 1970) In the absence of the law no one could trust any clothing label that it was not in the self-interest of the producer to specify correctly Hence these clothing labels, though incorrect, would not deceive many people

The law increases deceptive advertising in another way Consumers are unlikely to be legal experts Some, therefore, are likely to believe that certain forms of deceptive advertising are prevented by law when, in fact, they are not. It is not clear that broadening the definition of fraudulent advertising would solve this problem. The more the law protects against fraud, the more people think the law protects against fraud. Misinterpretation of the law's domain will exist, no matter how extensive that domain

Why, then, would the police power of the state ever be invoked against "deceptive" advertising, since it is quite possible that these laws increase rather than decrease deception in advertising? These laws can accomplish something They can—at a cost—make more information available to the consumer Consumer market power reduces deception by consumer distrust of any statement about which it is in the self-interest of producers to deceive As we have seen, for experience qualities this narrows considerably the information available to consumers from advertising Some people might deem it important that other information be made available from advertising, that consumers, for example, be able to determine the fabric content of their clothing Laws can achieve this objective at the price of both enforcement costs and costs to the consumer of the elimination of possible memorable sources of indirect information

These laws will be relevant almost exclusively to the advertising of experience qualities, since, even without laws, advertising for search qualities provides fairly complete information about the properties of the brand being advertised Deceptive advertising will be concentrated where the laws are concentrated—almost exclusively in the advertising of experience qualities There is evidence that supports this contention Though I find it unsatisfactory, let us use for the moment the Federal Trade Commission's criterion of what constitutes deceptive advertising Whatever its shortcomings, it has developed independently of the desire to generate data that would either support or reject hypotheses of economists For the first 6 months of 1965 the Federal Trade Commission found 58 advertisements deceptive about the quality of the product (as distinguished from its price) [16] All were advertisements about experience qualities (U S Federal Trade Commission 1970)

Clearly, then, there is some deceptive advertising The only empirical question is one of the magnitude of this phenomenon My analysis up to now has given little weight to deception in advertising What changes in this analysis are required if deception is an important rather than an unimportant part of advertising?

---

[16] The Federal Trade Commission did find a number of deceptive advertisements related to price Ordinarily the price of an article is known prior to purchase In all cases of deceptive advertising about price, however, the deceptive advertising was about price characteristics that would not be discovered prior to purchase When payment is postponed, the consumer need not know the price before the purchase This offers a firm an opportunity to place terms in the contract about which a large proportion of consumers would not be aware prior to purchase The Federal Trade Commission judged three of these cases in 1965 (U S Federal Trade Commission 1970) The other category of deceptive price advertising found by the Federal Trade Commission was the advertisement of a deceptively high regular price (six of these cases in the first 6 months of 1965) In purchasing the good, the consumer usually knows the purchase price but not the regular price In consequence this deception is not something that would be revealed to the consumer prior to purchase (The basis for the deception is that consumers often use regular price quotations to economize on search)

Surprisingly, this phenomenon would have little impact on the major conclusion of my analysis Look again at my demonstration that brands with the lowest P^*s (price/utility) have the greatest incentives to advertise It does not require intelligent consumer response to advertising—though it provides a basis for such intelligent response Consumers who actually believe paid-for endorsements are the victims of the most benign form of deception They are deceived into doing what they should do anyway. In consequence there should not be much difference between the behavior of those who respond to advertising because they are intelligent, and the behavior of the deceived Under these circumstances it is both exceedingly difficult and not very important to put the deception hypothesis to the test

This discussion of the deception hypothesis helps explain why economists have been so long deceived about the character of advertising It does not pay consumers to make very thoughtful decisions about advertising They can respond to advertising for the most ridiculous explicit reasons and still do what they would have done if they made the most careful judgments about their behavior

Whatever their explicit reasons, the consumers' ultimate reason for responding to advertising is their self-interest in so doing That is, it is no mere coincidence that thoughtful and unthoughtful judgments lead to the same behavior If it were not in consumer self-interest to respond to advertising, then consumers' sloppy thinking about advertising would cost enough that they would reform their ways

The learning required of consumers in this case is not very complicated Advertised endorsements have been used for a long time During this period the consumer, his parents, and his grandparents have consumed countless quantities of advertised and nonadvertised brands If consumers were losing out by trying advertised products, they would have had an enormous number of opportunities to discover this fact

Many economists have felt that other consumers think quite imprecisely about advertising—and well they might But this superficial observation had led economists, but not consumers, astray Economists have failed to see that consumers' response to advertising persists because of the underlying information role of advertising

Summary and Conclusion

This paper has attempted to show the way in which advertising as information operates I have contended that there is a basic difference in the character of the information conveyed by advertising for search and experience goods. For search goods advertising's information is direct. For experience goods the information conveyed is dominantly indirect—simply that the brand advertised. I have shown that this difference in the character of information leads to greater advertising expenditures for experience goods than search goods and greater marginal revenues to the consumer for search goods than experience goods. These implications were found satisfied by the data.

This analysis has not examined the totality of advertising's operation. I have not looked at the relationship between advertising and frequency of purchase of the advertised good—in contrast to brand. For example, food advertisements often contain recipes which provide information for the use of the good in question. Though this may be an important part of advertising's activity, it is irrelevant with respect to the set of implications examined in this article.

There is another possible function of advertising which I have not analyzed because I do not know how advertising's impact on a consumer's utility function, holding information constant. The change-in-taste idea cannot be effectively tested because no real theory about taste changes has been developed. Fortunately, one does not need such an idea to explain the major features of advertising behavior.

It should be perfectly obvious to anyone who has read this article that an enormous amount of work on advertising still remains to be done. On the theoretical side I have not provided a general equilibrium solution for the amount of advertising expenditures (in particular I have not solved for m in eq. [7]), on the empirical side I have not explained a substantial portion of the variance of advertising/sales ratios by industry. The focus of this paper has been neither to develop a complete theory of advertising nor to present a complete multiple regression analysis of advertising's empirical behavior. Rather, the emphasis in this paper has been to develop a theory of advertising that is both reasonable and productive. The productivity of this theory does not depend upon its completeness (as partial equilibrium analysis in general attests.) A theory is productive if it can generate implications that are consistent with real-world behavior. The theory developed in this article satisfies that requirement. That this theory makes sense out of a wide range of empirical phenomena suggests that it has caught the essence of advertising behavior, that further theoretical (or empirical) work would not vitiate our results.[17]

---

[17] For example, it has been suggested that the equilibrium that I envision cannot be stable, that all consumers would tend to shift to the better-advertised brand until that brand had all the customers. This contrary-to-fact result can easily be prevented by allowing consumer tastes and income to vary. Nonadvertised brands, then, would specialize in minority consumer taste. (While my analysis has for the most part assumed homogeneous consumers for simplicity, its conclusions do not depend upon this assumption.)

Appendix

Producers of equal efficiency have a choice of making a low-utility brand at high unit profits or a high-utility brand at lower unit profits It is reasonable to assume that the sales from customers who have no information about the brands' utilities will be the same for the two brands But consumers will tend to repeat the purchase of higher-utility brands In consequence, the expected sales of the high-utility brands will be greater In equilibrium, the total profits of equally efficient firms will be the same whether they make a high- or a low-utility product if firms of this level of efficiency continue to produce both This implies that the profit generated from the average individual making a purchase without information will be the same for both the high- and low-utility brands (Since both profits and number of customers with no information are assumed to be the same for both brands, the ratio of the two must be the same)

Assume, for the moment, that advertising merely increased the probability of a person's making an initial purchase of a brand without any impact on the conditional probability of a repeat purchase given an initial purchase Then high- and low-utility brands would have equal incentives to advertise, for the marginal probability of a repeat purchase would increase in the same proportion as the probability of an initial purchase

$$P(R) = P(I)\,P(R/I), \tag{A1}$$

where I = initial purchase, R = repeat purchase With $P(R/I)$ constant, $P(R)$ increases proportionately with $P(I)$ This implies that the average new customer generated by advertising would yield the same revenue to the firm as did the average old customer with zero information for both high- and low-utility brands Unless costs behaved in a very peculiar manner, this would imply that the profits generated by advertising were the same for the two brands, since the average old customer with zero information produces equal profits for the two brands

Why should this not be the state of affairs if advertising were simply informational? After purchasing (and using) a brand, the consumer does not require advertising as a clue to the hidden qualities of a good He can assess these hidden qualities directly Why, then, should advertising have any impact on the conditional probability of a repeat purchase?

A simple mechanism that produces this impact is the process of consumer memory The average consumer uses a lot of different products It costs him something to remember the brands of these products that he tries The cost is smaller the more familiar the name of the brand. Advertising makes brand names familiar We would expect, therefore, the consumer to remember a higher proportion of advertised brands that he has tried than unadvertised brands Indeed, there is evidence that this phenomenon does, in fact, operate The ratio of the recalled sales of advertised brands to recalled total sales of a product is higher than their actual market share (Sudman 1962)

This process should have an impact on the conditional probability of a repeat purchase. A necessary condition for a repeat purchase made deliberately is that the consumer remember the name of the brand of which he wishes to make a repeat purchase But if he remembers the name of the brand, he will not repeat his purchase by mistake For sufficiently low-utility brands, a consumer who remembers the name of the brand will have a lower conditional probability of a repeat purchase than if he did not remember the name of the brand For sufficiently high-utility brands, the reverse will be the case

For a high-utility brand, advertising will produce a greater percentage increase in the marginal probability of a repeat purchase than its impact on the probability of initial purchase, since both terms on the right in equation (A1) increase. Since we assume the impact of an equivalent dose of advertising on the probability of initial purchase to be the same for low- and high-utility brands, it follows that the producers of high-utility brands have a greater incentive to advertise than the producers of low-utility brands

References

Becker, Gary "Crime and Punishment An Economic Approach" *J.P.E.* 76, no 2 (March 1968) 169–217

Demsetz, Harold "The Nature of Equilibrium in Monopolistic Competition" *J P E* 67, no 1 (February 1959) 22–30

Dorfman, Robert, and Steiner, Peter "Optimal Advertising and Optimal Quality" *A E R* 44, no 5 (December 1954) 826–36

Doyle, P. "Advertising Expenditure and Consumer Demand" *Oxford Econ. Papers*, [illegible] 20, no 3 (November 1968) 394–414

Ferguson, James *The Advertising Rate Structure in the Daily Newspaper Industry* Engelwood Cliffs, N J Prentice-Hall, 1963

Gould, John "Diffusion Processes and Optimal Advertising Policy" In *Microeconomic Foundations of Employment and Inflation Theory*, edited by E Phelps et al New York Norton, 1970

Karlin, Samuel *A First Course in Stochastic Processes* New York Academic Press, 1966.

Life, vol 72 (1972)

Nelson, Phillip "Information and Consumer Behavior" *J P E* 78, no 2 (March/April 1970) 311–29

Schmalensee, Richard "A Note on Monopolistic Competition and Excess Capacity" *J P E* 80, no 3, pt 1 (May/June 1972) 586–91

Stigler, George "The Economics of Information," *J P E* 69, no 3 (June 1961) 213-25

——— "The Optimal Enforcement of Laws," *J P E* 78, no. 3 (May 1970) 526–36

Sudman, Seymour "On the Accuracy of Recording of Consumer Panels" Unpublished Ph D dissertation, Graduate School Bus , Univ Chicago, 1962

Telser, Lester, "Advertising and Competition" *J P E* 72, no 6 (December 1964) 537–62

U S , Bureau of the Census *Statistical Abstract of the United States, 1967* Washington: Government Printing Office, 1967

U S., Federal Trade Commission *Decisions Jan 1, 1965–June 30, 1965.* Washington, Government Printing Office, 1970.

Taxation and National Defense

Earl A Thompson

UCLA

Contrary to traditional models of national expenditures and taxation, national defense is not a consumer or a producer good Rather it is a good produced to protect the ownership of other goods from foreign takeover Using a model of international distributional equilibrium containing national defense and aggression decisions, a Pareto-optimal national tax structure is derived The entire U S tax structure is then shown to be roughly equivalent to the optimal structure This empirical result indicates that the U S political system has somehow been able to create a roughly efficient tax structure in spite of the systematic misguidance of economists

In the traditional theory of national taxation, the government provides only collective consumer or producer goods Because governmental provision of such goods does not alter any of the familiar necessary conditions for the efficient utilization of the resources remaining for private use, efficient taxation when the private-goods sector is perfectly competitive is achievable only with lump-sum taxes (such as head taxes, land taxes, or equal, ad valorem taxes on all consumer benefits) Any tax other than a lump-sum tax yields violations of the familiar necessary conditions for achieving Pareto optimality given the resources remaining for private use

In the real world, lump-sum taxes have been consistently rejected despite the *obiter dictum* of the traditional theory of taxation

This paper begins with a development of a theory of efficient national taxation in a world in which each government provides collective defense of the property of its citizens While the traditional model is frequently

This research was supported by a Lilly Foundation grant for the study of property rights at UCLA The author benefited from comments on previous drafts made by Armen Alchian, Arnold Harberger, Louis Makowski, Bob Rooney, and a referee of this *Journal*.

alleged to apply to collective defense, it is an error to describe collective defense as a collective consumer or producer good. A tank is neither a consumer nor a producer good. In the descriptively more accurate model of collective defense developed in Section I below, the private accumulation of certain kinds of capital creates an extra defense burden for the country protecting that capital. Efficient taxes are then not lump sum taxes; efficient taxes discriminate against the accumulation of capital that is coveted by potential foreign aggressors.

Section II shows that the optimal capital tax is achieved by the use of a simple income tax complemented by (1) realistic depreciation allowances, (2) tax write-offs for charity and abnormal, noncosmetic, medical expenses, and (3) a theoretically specified, positive (*a*) percentage depletion allowance to natural resource owners, (*b*) degree of progression in the income tax rates, (*c*) minimum income exemption for each individual and his dependents on his income tax, (*d*) corporate-profits tax, (*e*) excise tax on each consumer durable, and (*f*) tax break on capital gains.

In contrast, in the traditional economic theories of government expenditures and taxation all of these special features of the income tax—and the income tax itself—are inconsistent with a Pareto optimum.[1]

Section II also presents rough-and-ready quantitative estimates of the six optimal rates in consideration 3 above. The results are compared with actual U.S. rates and indicate a striking degree of overall efficiency of the U.S. tax structure. Our recommendations for policy changes have relatively minor effects on economic incentives.

The results raise the question of what in the U.S. political system has permitted the evolution of such an efficient tax structure in spite of the fact that the only existing tax theories deprecate every major feature of the structure.

I. The Basic Model

The Environment

We shall employ a capital model with discrete time, an infinite horizon, no joint production, and perfect knowledge about the environment.[2]

---

[1] Samuelson (1964) has attempted to provide some kind of rationalization of realistic depreciation allowances in an apparently traditional environment, but we shall find a simple economic error in his exercise. Also, some authors have assumed that an income tax does not generate a "double taxation of savings" and an inefficiently discriminating tax on future (relative to present) consumption in the traditional model. Since the double tax is only on the interest from savings in the traditional model and there may be a zero interest rate, these authors can be defended, but only in a special case.

[2] The central results in this paper have also been established for a mathematically less familiar model containing continuous-time and Marshallian joint production, although dropping this assumption requires a substitute assumption, as noted at the

We assume that each of the N^α individuals in country α has a utility function defined over all feasible sequences of his consumption benefits,

$$U^i(B_1^i, B_2^i, \quad), \qquad i = 1, 2, \quad , N^\alpha,$$

where $U^i(\cdot)$ is a monotone increasing, differentiable, strictly quasi concave function [3] Aggregate consumption benefit over these individuals during the tth period is given by the differentiable, quasi concave production function

$$\sum_{i=1}^{N^\alpha} B_t^i = C^t(K_{1Ct}, K_{2Ct}, \quad , K_{MCt}), \qquad t = 1, 2, \quad , \tag{1}$$

where K_{kCt} represents capital of the kth kind ($k = 1, \quad , M$) devoted to the production of consumption benefits at time t Aggregate capital in country α in each future period is the result of devoting capital in the preceding period to its production or acquisition so that

$$K_{kt+1} = I^{kt}(K_{1kt}, \quad , K_{Mkt}), \qquad t = 1, 2, \quad , \tag{2}$$

where $I^{kt}(\ )$ is differentiable and quasi concave Foreign aggression is treated as a form of investment, the capital obtained from a current act of aggression being unavailable for consumption until the following period

National defense effort at time t is

$$D_t = G^t(K_{1Gt}, \quad , K_{MGt}), \tag{3}$$

where $G^t(\ )$ is also differentiable and quasi concave The aggregate capital stock of kind k in any period is the sum of the amounts of capital used in the above activities plus the amount taken by foreign aggressors, K_{kAt} That is,

$$K_{kt} = K_{kCt} + \sum_{y=1}^{M} K_{kyt} + K_{kGt} + K_{kAt} \tag{4}$$

---

end of this section of the paper In the case of a finite horizon, several of our tax-equivalence theorems hold in only an approximate sense The perfect-information assumption can be substantially relaxed and is made largely to simplify the discussion

[3] The "strict quasi concavity" of a function here means that if $f(x) = f(x')$, $x \neq x'$, then

$$f[ax + (1 - a)x'] - f(x) > \delta \sup_j (x_j - x_j') f_j$$

for some $\delta > 0$ where $1 > a > 0$ and $x = x_1, x_2,$ This may also be termed "asymptotically strict quasi concavity " It assures the absence of infinite quantities in maximizing f over all x subject only to linear equalities in x with positive coefficients This type of quasi concavity, as well as the standard quasi concavity assumptions stated below [$f(x)$ is quasi concave if $f(x) = f(x')$, $x \neq x'$ implies that $f[ax + (1 - a)x'] - f(x) \geq 0$], also are used to insure the existence of a competitive equilibrium. The differentiability assumptions in this paper, together with the absence of traditional nonnegativity assumptions, are made to facilitate the mathematical argument, since it is obvious that technical nonsubstitutabilities do not disturb the optimality of incentives in our efficient tax system.

Relations (1)–(4) insure the absence of joint production and collective goods in that they state that no particular unit of a capital good serves several functions simultaneously, such as producing consumption goods and producing itself in the following period.

Similar relations hold for each country.

We assume that the distribution of capital between countries is an equilibrium distribution, meaning that each of the countries has rationally decided (rational in the Paretian sense) which property to claim and defend in each period and that the decisions are mutually consistent given the world's aggregate stocks of capital. The equilibrium initial capital stock of country α is given by

$$K_{k1} = K^{*}_{k1}, \qquad k = 1, 2, \ldots, M \tag{5}$$

We assume that each country knows the rational strategies of the others so that no actual aggression occurs in determining this equilibrium distribution of capital between countries. In equilibrium, the country possessing a unit of capital in a given period is the country that has made a prior commitment to impose on any other country attempting to acquire the capital damages which are at least as great as the value of the capital to that country.[4] Since the only way to subvert the prior commitment of another country is to take control over the entire country, all foreign aggression is all or nothing. So if the net return from aggression is ever positive, it is greatest for $K_{kAt} = K_{kt}$. Hence, we can write the profit to aggression against country α for a particular aggressor as

$$\pi_{At} = A(K_{1t}, K_{2t}, \ldots, K_{Mt}) - C(D_t), \tag{6}$$

where $A(\ )$ is the aggressor's evaluation function of the assets he acquires and $C(D_t)$ is his corresponding cost of the aggression, which we assume to be a monotonic increasing function. The rational foreign aggressor acquires no K_{kt} if and only if the profit to the aggression is never positive.

---

[4] It may be of interest to note that such commitments are generally impossible to make in a pure democracy, as the voters can always vote against a war by voting down war appropriations. Such a government is sure to lose, bit by bit if not all at once, all of its transferable capital to a nondemocratic foreign aggressor that can make commitments to fight wars over property at war costs to both parties which are greater than the value of the property at stake. (The reason such commitments are rational is that once they are made, the democratic country rationally surrenders its capital so that fighting the war is unnecessary.) However, in a constitutional democracy, where certain government policies are not subject (except at great cost) to future voter disapproval, the constitution, by giving proper incentives to the government leaders and by allowing them to command war resources without voter approval, may effect the necessary war commitment. Constitutionally granted wartime finance policies such as the draft, debt financing and government currency creation, and price controls are thus a necessary part of our wartime financial structure. An important implication of this necessary, confiscatory, wartime financing is that there is an insufficient accumulation of war relevant capital during peacetime. This is the economic basis of the classical "national defense argument" for peacetime subsidization of certain domestic industries (see Thompson [1968] for an elaboration and empirical application). In the formal model above, such subsidies appear as government purchases of capital used to produce national defense.

Equilibrium also implies that each country always retains its rationally produced capital This is true because if a subsequent unit of capital were not successfully defended, it would have been better for the country to consume the capital which was devoted to its production This follows from the monotonicity of utility functions and the consumability of capital expressed in relations (2)–(5). In equilibrium, then, α makes D_t just high enough in each period that for each potential aggressor, $\pi_{At} \leq 0$ for all $K_{kAt} > 0$ That is, it makes D_t just high enough that the solution value of K_{kAt} is equal to zero for all potential aggressors From (6), this level of D_t obviously depends upon $K_{1t}, \quad, K_{Mt}$, which determines the return to foreign aggression Hence, we set D_t equal to $D^t(K_{1t}, \quad, K_{Mt})$, the defense requirement of the country at time t—the level of defense required to dissuade all potential aggressors Also, since there are no foreign aggression activities in equilibrium, the only output of the "government" is D_t Hence, equations (3) and (4) are written

$$D^t(K_{1t}, \quad, K_{Mt}) = G^t(K_{1Gt}, \quad, K_{MGt}), \tag{3'}$$

$$K_{kt} = K_{kCt} + \sum_y K_{kyt} + K_{kGt} \tag{4'}$$

Conditions for Pareto Optimality

Maximizing $U^1(B_1^1, B_2^1, \quad)$ subject to $U^j(B_1^j, B_2^j, \quad) = U^{j*}$, $j \geq 2$, and equations (1), (2), (3′), (4′), and (5), we find that necessary for a Pareto optimum in our environment is that the allocation of resources (i e , K_{kCt}, K_{kyt}, K_{kGt}, and B_t^i for all k, y, t, and i) satisfies, in addition to the constraint equations above, the following marginal equalities

$$\frac{C_k^t}{C_y^t} = \frac{I_k^{zt}}{I_y^{zt}} = \frac{G_k^t}{G_y^t} \qquad \text{for all } t, k, y, \text{ and } z, \tag{7}$$

and

$$\frac{\partial U^i/\partial B_t}{\partial U^i/\partial B_{t+1}} = \frac{C_y^{t+1} I_k^{yt}}{C_k^t}, \left(1 - \frac{D_y^{t+1}}{G_y^{t+1}}\right) \qquad \text{for all } t, k, y, \text{ and } i, \tag{8}$$

where subscripts on function symbols indicate partial derivatives of the function with respect to capital of the type specified by the subscript

Equation (7) states the familiar condition that in an optimum, different kinds of capital are allocated between sectors so that their relative marginal productivities are equal whatever they produce Equation (8) states that in an optimum, the marginal rate of time preference of B_t over B_{t+1} is less than the familiar marginal rate of time productivity by a percentage equal to the increase in defense requirement caused by the capital which is produced to create the extra B_{t+1} relative to the defense productivity of this capital.

Competitive Equilibrium

We now give each individual in country α an initial endowment of capital $(K_{11}^{i*}, \ldots, K_{M1}^{i*})$ such that

$$\sum_i K_{k1}^{i*} = K_{k1}^* \qquad \text{for all } k \tag{9}$$

We also give each of these individuals a set of quasi concave production functions for each period which read

$$\begin{aligned} C^{ti} &= C^{ti}(K_{1Ct}^i, \ldots, K_{MCt}^i) && \text{and} \\ I^{kti} &= I^{kti}(K_{1kt}^i, \ldots, K_{Mkt}^i) && \text{for every } k \end{aligned} \tag{10}$$

The aggregate functions described in (1) and (2) must then be derived by maximizing aggregate output for given aggregates of inputs devoted to the production of the output. That is,

$$C^t(K_{1Ct}, \ldots, K_{MCt}) = \max \sum_i C^{ti}(K_{1Ct}^i, \ldots, K_{MCt}^i)$$

subject to

$$\sum_i K_{yCt}^i = K_{yCt},$$

and (11)

$$I^{kt}(K_{1kt}, \ldots, K_{Mkt}) = \max \sum_i I^{kti}(K_{1kt}^i, \ldots, K_{Mkt}^i)$$

subject to

$$\sum_i K_{ykt}^i = K_{ykt}, \qquad \text{all } k, y, \text{ and } t$$

We assume that each individual may economically participate somewhat in the production of each output (which will imply nonincreasing returns to scale in the individual functions), so that the above maximizations obviously occur when and only when

$$\begin{aligned} C_k^{ti} &= C_k^{tj} \; (= C_k^t), \\ I_k^{yti} &= I_k^{ytj} \; (= I_k^{yt}) && \text{for all } k, i, j, \text{ and } t \end{aligned} \tag{12}$$

Hence, if the equilibrium in the economy satisfies (12), it generates the aggregate production functions in (1) and (2) given the constraints in (11).

We now introduce prices. Our prices are all initial-period, unit-of-account prices; that is, they describe the amount of wealth one must currently surrender in order to obtain delivery of a good at a specified date. To obtain such prices from prices that would rule in actual transactions in future periods, a suitable discount must be applied to the price in the future to reflect the value of early payment in the form of initial wealth. The price of capital of type k delivered in period t is written P_{kt}, and the price of consumption goods delivered in period t is written P_t. An individual is also taxed an amount whose present cost is given by T^i

Each individual is assumed to choose B_t^i, K_{kCt}^i, and K_{kyt} so as to maximize $U^i\,(B_1^i, B_2^i, \quad)$ subject to the production functions in (10) and his budget,

$$T^i + \sum_t P_t B_t^i = \sum_k P_{k1} K_{k1}^{i*} + \sum_t P_t C^{ti} + \sum_{ty}\sum P_{yt+1} I^{yti} \qquad (13)$$
$$- \sum \sum_{tyk} \sum P_{kt}(K_{kCt}^i + K_{kyt}^i)$$

The solutions represent a competitive equilibrium when prices are set[5] so that, for all t and k,

$$\sum_i Z_t^i = \sum_i C^{ti}, \qquad \sum_i (K_{kC1}^i + \sum_y K_{ky1}^i) = K_{k1}^*,$$

and (14)

$$\sum_i (K_{kCt+1}^i + \sum_y K_{kyt+1}^i) = \sum_i I^{kti}$$

The Case of Lump-Sum Taxation

When taxes are lump sum so that they do not vary with the individual's behavior, the individual utility-maximizing choices are seen to satisfy the following marginal equalities

$$\frac{C_k^{ti}}{C_y^{ti}} = \frac{I_k^{zt}}{I_y^{zt}} = \frac{P_{kt}}{P_{yt}}$$

and

$$\frac{\partial U^i/\partial B_t}{\partial U^i/\partial B_{t+1}} = \frac{C_y^{t+1} I_k^{yt}}{C_k^t} = \frac{P_t}{P_{t+1}}, \qquad \text{all } k, y, z, t, \text{ and } i$$

These conditions are inconsistent with the condition for Pareto optimality in (8) except when $D_y^{t+1} = 0$ for every y and t, which is the implausible special case in which the returns to aggression by the marginal foreign aggressors are never affected by the size of the victim's capital stock

The Pareto Optimality of a Competitive Equilibrium with Certain Capital Taxes

We now assume that

$$T^i = \sum_{tk}\sum a_{kt} P_{kt} K_{kt}^i + T^{oi}, \qquad (15)$$

---

[5] We assume that

$$\liminf_{t \to \infty} \left(\sum_{ty}\sum P_{yt+1} I^{yt} - \sum \sum_{tyk} \sum P_{kt} K_{kyt} - T^t\right)$$

exists to assure finite solutions Our prices are therefore "Malinvaud prices" (Malinvaud 1953) where a_{kt} is a constant present tax rate on capital of type k at date t and the T^{oi} is a lump-sum tax or subsidy to the individual set so that $\sum_i T^i$ satisfies the government's budget condition,

$$\sum_i T^i = \sum_{tk} \sum P_{kt}(1 + a_{kt}) K_{kGt} \tag{16}$$

Equation (16) reflects the fact that the capital tax is levied on sellers of capital rather than buyers and that prices are the net prices to sellers We assume that the government minimizes costs using fixed-factor prices so that

$$\frac{P_{kt}(1 + a_{kt})}{P_{yt}(1 + a_{yt})} = \frac{G_k^t}{G_y^t} \tag{17}$$

Maximizing $U^i(\ )$ subject to (10), (13), and (15) yields the following marginal conditions

$$C_k^{ti} = P_{kt}(1 + a_{kt})/P_t \quad \text{and} \quad I_k^{yti} = P_{kt}(1 + a_{kt})/P_{yt+1}, \tag{18}$$

$$\frac{C_k^{ti}}{C_y^{ti}} = \frac{I_k^{zti}}{I_y^{zti}} = \frac{G_k^t}{G_y^t} = \frac{P_{kt}(1 + a_{kt})}{P_{yt}(1 + a_{yt})}, \tag{19}$$

and

$$\frac{\partial U^i/B_t}{\partial U^i/B_{t+1}} = \frac{C_y^{t+1} I_k^{yt}}{C_k^t} \frac{1}{1 + a_{yt+1}} = \frac{P_t}{P_{t+1}}, \qquad \text{all } k, y, z, t, \text{ and } i \tag{20}$$

Equation (17) satisfies the conditions in (12) so that (11) holds We now need only set

$$a_{yt+1} = a^o_{yt+1} = \frac{D_y^{t+1}}{G_y^{t+1} - D_y^{t+1}} \tag{21}$$

in order for (19) and (20), together with (9), (11), (14), (3′), and (4′) to represent the same equation set as (7) and (8) together with (1), (2), (3′), (4′), and (5)—in order for any competitive equilibrium with such capital taxes to be a Pareto optimum [6]

---

[6] Since we have not precluded decreasing marginal costs of defense with respect to the protected capital, there may be several allocations satisfying (7) and (8) for given utility levels of $N^a - 1$ individuals And some of these allocations may be Pareto nonoptimal In such cases, however, we assume that the government picks tax rates (a_{kt} and T^{oi}) which correspond to their shadow values in a global Pareto optimum. That such taxes succeed in inducing global Pareto optima despite the possible economies of scale in protecting capital is easily proved First note that our quasi concavity, nonincreasing returns and bounded-wealth assumptions insure that all solutions to (19) and (20)—given (9), (11), (14), (3′), and (4′)—are privately optimal. It follows that we need only pick the levels of a_{kt} and T^{oi} that correspond to their shadow equivalents in any specified Pareto optimum in order to have the privately selected allocations of the equilibrium coincide with the Pareto optimum For if this were not the case, then one allocation satisfying (19) and (20)—given (9), (11), (14), (3′), and (4′)—would make someone worse off than would another such allocation

Note that no particular tax rate on capital in the initial period is implied by optimal capital taxes. This is reasonable because such capital has already been produced so that taxing it is equivalent to applying a lump-sum tax. Nevertheless, we oftentimes below apply the harmless procedure of applying the optimal tax rate on future capital to capital in the initial period. Note also that the optimal capital tax is equivalently a tax on the value of the capital output, I^{kt}

It is easy to show, using (18), allowing the optimal capital tax rate to apply in period 1, and assuming linear homogeneity of $D^t(\ )$ and $G^t(\ )$ with respect to their respective arguments, that $T^{ot} = 0$ so that optimal capital taxes alone are just sufficient to finance government expenditures. While we do not maintain these homogeneity assumptions in the paper, the result indicates, to the extent that the homogeneities are roughly plausible, a relatively minor role for lump-sum taxation or subsidization in a world employing an optimal capital tax

Specification of the Marginal Aggressors' Marginal Profit Functions

Since the marginal aggressors' profits are kept at zero by the potential victim's defense effort, we have, differentiating (6),

$$\frac{\partial A}{\partial K_{xt}} = A_{xt} = \frac{dC(D^t)}{dD^t} D^t_x \tag{25}$$

Hence,

$$\frac{A_{xt}}{A_{yt}} = \frac{D^t_x}{D^t_y} \qquad \text{if } A_{yt} > 0 \tag{26}$$

We assume that for some subset of $(k) = (1, 2, \quad, M)$, written $a(k) = (1, 2, \quad, M_a)$, $A_{kt} > 0$ We call any kind of capital in this subset a part of the country's "coveted capital " For the rest of the capital stock, $A_{kt} = 0$

Equations (21) and (25) tell us that if $A_{xt+1} = 0$, $a^*_{xt+1} = 0$ That is, if a particular kind of capital is not part of the country's coveted capital, the optimal tax on the capital is zero Now we assume that for all x and y in $a(k)$,

$$\frac{A_{xt}}{A_{yt}} = \frac{P_{xt}}{P_{yt}} \tag{27}$$

That is, the relative marginal values which foreign aggressors place on different kinds of the country's coveted capital are equal to the corresponding relative values to the defending country There are several reasons that this is not a strictly justifiable assumption. It does, however, serve to maintain reasonable orders of magnitude. A jet plane is a lot more valuable than a light bulb, to the aggressor as well as the defender. From (26) and (27),

$$\frac{P_{xt}}{P_{yt}} = \frac{D^t_x}{D^t_y} \qquad \text{for all } x, y \in a(k) \tag{28}$$

Hence, from (28), (19), and (21), optimal capital taxes are nondiscriminatory, that is,

$$a^o_{xt} = a^o_{yt} = a^o_t \qquad \text{for all } x, y \in a(k) \tag{29}$$

Problems in Implementation

We have as yet produced no model specific enough to indicate which types of capital comprise a country's coveted capital stock Also, since it is practically very costly to tax the value of capital in every period when there are not transactions in the capital during every period, a problem arises as to how one can create, if possible, a tax system which levies only *on transactions but which is still equivalent in effects to the idealized* system of optimal capital taxation described above These problems of implementation are the subject of Section II

A Possible Generalization

Admitting Marshallian joint production of consumption and investment goods would open up the possibility that some units of produced capital would be optimally surrendered in the future to a potential foreign aggressor To have determinate units of such capital, we would, in effect, have to introduce different defense costs for different units of equally valuable capital, thus violating (28) and giving rise to discriminatory taxes While these cases of "surrenderable capital" are excluded from our formal analysis below, an informal discussion is provided at the end of Section II

II. Achieving an Optimal Capital Tax

The Transaction Structure and Income Taxes

We now allow our economy to have an explicit transaction structure—a particular set of trades between individuals which achieves the optimal competitive equilibrium described above Suppose each producer sells his entire output, purchasing all of his inputs all over again for his production in the following period Then there would be no difficulty in implementing the optimal capital tax A simple income tax, a tax on all producer sales, with tax exemptions granted for sales of the outputs of noncoveted capital, would obviously be sufficient to produce an equivalent to the optimal capital tax. However, this supposition is far from realistic, in our model, producers may retain some of their capital output for their own future use To acquire an equivalence between an income tax and the optimal capital tax then requires amendments to the simple income tax besides exemptions for sales of outputs from noncoveted capital The income tax on the *k*th kind of capital of individual i is given by

$$T_b^{ki} = \sum_t b_{kt}(P_t C_k^t K_{kt}^{im} - X_{kt}^i), \tag{30}$$

where b_{kt} is the income tax rate, X_{kt}^i represents deductions from the tax base, and K_{kt}^{im} represents the capital that i uses to produce goods for the market in period t By definition,

$$K_{kt}^i = K_{kt}^{im} + K_{kt}^{ii}, \tag{31}$$

where K_{kt}^{ii} is the capital that individual i uses to produce goods which are not sold in the market The optimal income tax exists when, for each k and i, b_{kt} and X_{kt}^i take on values that make (30) equivalent to (15) and (21), or,

$$\sum_t b_{kt}(P_t C_k^t K_{kt}^{im} - X_{kt}^i) = \sum_t a_t^o P_{kt} K_{kt}^i + T^{oi} \qquad \text{for all } K_{kt}^i, \tag{32}$$

or

$$dT_b^{ki}/dK_{kt}^i = a_{kt}^o P_{kt} \tag{33}$$

Solutions to (32) and (33) are written (b_{kt}^o, X_{kt}^{io})

Note that income taxes have an independence property expressed as

$$\sum_k T_b^{ki} = T_b^i \tag{34}$$

Noncoveted Capital

Capital may generate benefits which apply to only certain, specified individuals within a country—such capital is "people specific " One's psychic capital—for example, acquired ability to appreciate nature and various activities, stock of pleasant memories, acquired ability to entertain himself—is part of the country's people-specific capital stock The rest of a country's "people-specific capital" is "friendship capital"—where a specified individual (e.g., a husband) can command the services (e g , good cooking) of another (e.g , a wife) because the other either feels a sense of gratitude or indebtedness to the specified individual or has confidence that the specified individual will reciprocate in the future without an explicit agreement, the cost of which would preclude such favors from having a positive net value to the recipient. People-specific capital is of no value to individuals outside the country and is not part of a country's coveted capital stock.[7] Since the returns on people-specific capital do not generate market transactions so that $K_{kt}^{im} = 0$ for people-specific capital, an income tax operates in an optimal fashion by not taxing people-specific capital

While returns from friendship capital frequently come in the form of a monetary payment in the cases of gift and charity income, such income is, appropriately, substantially disregarded for tax purposes in the United States [8] However, large gifts (including inheritances) frequently only partially represent a payment for friendship services, with another part representing a payment for services that a foreign aggressor could acquire Large gifts and inheritances should therefore be taxed to some degree, with the tax rate increasing with the size of the gift or inheritance and never exceeding the income tax rate We do observe such taxes, although we have no way of estimating optimal rate schedules to compare to the actual schedules

When an aggressor takes over a country, he does not benefit from human capital to the extent that such capital is necessary for the "subsistence" of the individual An individual is below "subsistence" if he would sooner die in rebellion than pay the taxes of the successful foreign aggressor One's subsistence is comprised of normal support for his family and any abnormal, noncosmetic, family medical expenses Hence, one's human capital below that required for his family's subsistence is not part of his country's coveted capital stock As a result, an optimal income tax has an exemption, X_{ht}, on incomes from human capital required for normal family subsistence of market goods and for abnormal, noncosmetic medical expenses The observed U S income tax exemption of about $750 per person appears to approximate the normal subsistence level fairly well, while the observed write-off for abnormal, noncosmetic medical expenses corresponds nicely to the treatment of these expenses under an optimal income tax

It is likely that a modern foreign aggressor would support those citizens of an acquired country who were unable to produce their own subsistence Assuming this to be the case, any increase in the assistance requirement of below-subsistence individuals in a defending country has the same external social product as any other decrease in the coveted capital stock of equal value. The decrease in the coveted capital stock in this case appears as an increase in expenditures on charity Hence, a receipt of charity is not only a return on noncoveted capital, it is also a reward to individuals who produce less than their subsistence for reducing their country's stock of coveted capital So expenditures on charity should be treated for income tax purposes as any other expenditure which reduces the capital stock of the country In fact, charitable expenditures in the United States are treated substantially the same as business expenditures for income tax purposes [9]

---

[7] It should be pointed out that our assertion that people-specific capital is noncoveted rests on an assumption that a successful foreign aggressor cannot substantially switch this kind of capital into the production of benefits which are not people specific Thus, it is assumed that the human capital that an individual devotes to producing benefits for himself (i.e., producing "leisure") cannot be substantially converted to the production of goods for the aggressor Our foreign aggressors therefore do not make slaves out of their victims, they merely tax them to subsistence This implication appears to be fairly realistic The behavioral basis of it appears to be that an individual acquires certain work-leisure habits, which cannot be substantially broken at any reasonable cost by the foreign aggressors

[8] A gift which does not represent a payment for a particular service but a mere redistribution of wealth based upon directly interdependent utilities, while not contained in our formal model, is clearly a mere transfer and should not be taxed

Certain consumer durables such as furniture, portraits, trophies, and certain antiques are also part of the country's noncoveted capital stock While there are no taxes on these goods as $K_{ct}^{im} = 0$, consumer durables which are part of the coveted capital stock are discussed below

Cash, that is, paper currency, may exist in the economy with explicit transactions as an intermediate asset which allows for the achievement of the no-transactions-cost economy described in Section I However, with costless currency creation, a successful foreign aggressor can also costlessly create any feasible level of real cash balances by altering the rate of growth of the currency supply Hence, the level of real cash balances used by the defending country is irrelevant to the aggressors and not part of a country's coveted capital stock

The definitions of capital which Marshall (1925), Knight (1935), and Smith (1937) inferred from discussions of men of affairs excluded people-specific capital, subsistence-producing capital, certain personal consumer durables, and paper money Using our theory, we can rationalize the exclusion of these forms of capital from the concept of capital used by men of affairs by arguing that these men are only discussing coveted capital because it is the only capital which should be taxed The only remaining types of private-good capital in the Smith-Marshall-Knight taxonomies—the types comprising the capital stock as seen by men of affairs—are natural resources, produced producer durables, human capital above subsistence requirements, and the remaining consumer durables These forms will be assumed to comprise the country's coveted capital stock, and the achievement of an optimal tax on each of these forms of capital is examined separately below

---

[9] Note, however, that a standard-type competitive model such as that developed above applies only to cases of private-good charity—cases in which each individual enjoys *his* giving charity rather than enjoying the beneficiary's receipt of the charity (i.e , in which there are no external economies in the giving of charity) and in which an individual can costlessly avoid the knowledge of another's suffering (i e , in which there are no external diseconomies in the creation of charity-inducing attributes) Including both of these excluded realistic cases would probably have little total effect on the tax treatment of charity because their separate effects are substantially cancelling

Producer Durables Which Are Originally Produced for Sale

No tax on the gross income from capital which is not sold in every period is equivalent to the optimal tax The derivative of taxes with respect to K^t_{kt} is positive in (33), and, in view of (31), is zero in (30) for fixed K^{tm}_{kt} if $X^t_{kt} \equiv 0$ So, for optimal taxation, $X^t_{kt} \not\equiv 0$ To achieve the effects of an optimal capital tax, we wish to consider a tax on *k*'s gross income reduced by an estimate of its depreciation

Depreciation, however, is generally estimable only for a good which is originally produced for sale, meaning that whenever a good other than itself is used to produce it, it is marketed Without loss of generality, we assume that when a producer good that is originally produced for sale (labelled *e* for equipment) is not used to produce current marketed output, it is used to produce only itself in the following period That is, $K^{tt}_{et} = K^{tt}_{eet}$

Let[illegible] X_{et} represent realistic depreciation of the original purchased capital in terms of current consumer goods, (30) becomes

$$T_b^{et} = \sum_t b_{et}[P_t C_e^t (K^t_{et} - K^{tt}_{eet}) - P_{et+1}(K^t_{et} - I_e^{et} K^{tt}_{eet})]$$

Using (18) this becomes

$$T_b^{et} = \sum_t b_{et}(P_t C_e^t - P_{et+1}) K^t_{et} \tag{35}$$

Hence, to obtain an optimal net income tax on *e* in time *t*, in view of (32) and (35), we need only pick a b_{et} which satisfies

$$b_{et} = \frac{a_t^o P_{et}}{P_t C_e^t - P_{et+1}} \tag{36}$$

Using (18), this optimal income tax rate becomes

$$b^o_{et} = \frac{a^o_t}{1 + a^o_t} \; \frac{I_e^{et}}{I_e^{et} - 1} \tag{37}$$

From (20) and (18) we can see that when there is a stationary solution such that $a^o_t = a^o_{t+1}$ and $C^t_k = C^{t+1}_k$,

$$b^o_{et} = \frac{a^o_t (1 + \rho_t)}{(1 + \rho_t)(1 + a^o_t) - 1}, \tag{38}$$

where $1 + \rho_t = (\partial U/\partial B_t)/\partial U/\partial B_{t+1})$, the marginal rate of time preference.

Thus, the use of realistic, physical depreciation allowances converts the inefficient income tax into an efficient one under an appropriate income tax rate[10]

---

[10] In the real world, depreciation allowances are typically granted on a fixed schedule for a particular kind of capital good regardless of how the good actually wears out But, also in the real world, depreciation typically takes the form of Marshallian joint pro-

The only other existing rationalization of the realistic depreciation write-off this author has seen is by Samuelson (1964) Samuelson shows that a write-off on income taxes of all forms of interest expense will create a tax structure with no effect on prices only if the tax on income includes a write-down for realistic depreciation However, no tax system in existence subsidizes interest expenses and taxes net receipts so as to have offsetting revenue and incentive effects Such a system would be a lot of trouble for no purpose The tax system Samuelson claims to represent is a United States-like system, a system containing a tax on interest income as well as a corresponding write-off of interest expenses But in such a system there is no net tax or subsidy on borrowing or lending—nor should there be in our optimality model because borrowing represents a mere redistribution of purchasing power rather than the creation of any real asset Samuelson erred by failing to allow the after-tax contractual rate of interest to rise to reflect the equal shifts up in the interest-demand price for loans and the interest-supply price of loans resulting, respectively, from the tax write-off of interest expenses and the tax on interest income Once gross market rates of interest are raised to reflect the tax on interest income, the reduction of the gross market rate by applying the tax write-off on interest expenses (or income) to obtain the after-tax borrowing (or lending) rate relevant for discounting only serves to pull the discount rate back down to the original real interest rate for the original allocation of real resources And with no reduction in the discount rate in Samuelson's model, no positive income tax rate satisfies Samuelson's price invariance condition, whether or not there are depreciation allowances

Natural Resources

Natural resources (i e , minerals and oil and gas) pose a different problem in the lack of transactions to correspond to each act of production Like producer-goods which are purchased upon their original creation, natural resources are accumulated by the owner without any corresponding transaction But unlike such producer goods, there can be no transaction

---

duction of marketed output and future capital Under such depreciation, the optimal tax on a new investment under stationary conditions is a single tax on the present value of the capital in each future period resulting from the investment, while an income tax without a depreciation allowance is a tax only on the initial capital value It is easy to see by an argument similar to that used above that a realistic depreciation allowance converts the income tax into a tax on the present value of the future capital values implied by an investment and thus leads to an optimal choice of investments in a world in which actual depreciation takes the form of Marshallian joint production of marketed output and future capital for the firm

There are real-world cases of physical depreciation of producer durables which have been produced for sale in which the depreciation does not take the form of Marshallian joint production. This appears for originally produced timber, wine, and various agricultural products Here, we observe in the United States, the variable depreciation allowances of the kind which our formal model describes, these are frequently called "cost depletion allowances " corresponding to original creation for a natural resource. Because tax collectors cannot be assumed to know the value of the natural resource at any given date in the past, they cannot be assumed to know the change in its value over time However, once a natural resource is utilized by converting it into some other good by an act of withdrawal from nature, there is, we assume, always a sale of the withdrawn resource. Hence, the obvious method of achieving the effects of a tax on the accumulation of natural resources is to apply taxes at withdrawal in a way which subsidizes early withdrawal The ordinary income tax does not do what it may appear to—tax early withdrawal—because the profit from withdrawal increases with the rate of interest so that delaying withdrawal merely increases future taxes by the rate of interest and has no tax-saving or tax-increasing effect in a world with a constant income tax rate Obviously, a subsidy to net withdrawal income would not encourage early withdrawal We can specify a tax or subsidy on transactions which will encourage early withdrawal only after specifying some special, technological features of the natural resource industry

In particular, (*a*) in producing a natural resource for next year, the only input one uses is the same resource in the current year, and (*b*) the amount of the natural resource produced is identical to the amount of the resource devoted to its production. That is,

$$K^{i}_{nt+1} \equiv I^{nti} \equiv K^{ii}_{nnt}, \tag{39}$$

where *n* is the natural resource From this and (18), in an optimum,

$$P_{nt}(1 + a_{nt}) = P_{nt+1} = P_t C^i_n \tag{40}$$

Similarly, the same physical units of natural resources become consumable once they are withdrawn so that, letting Q^i_t be *i*'s withdrawal of natural resources in time *t*, $P_{Qt}K^i_{nQt}$ represents the consumption value of the withdrawn resource It follows that

$$P_{Qt} - P_{nt}(1 + a_{nt}) = P_{Qt} - P_{nt+1}$$

is the direct withdrawal cost per unit of the resource assuming linear homogeneous withdrawal functions [11] Thus,

$$P_t C^i_n = \frac{P_{Qt}\,\partial K_{nQt}}{\partial K_{nCt}} - (P_{Qt} - P_{nt+1}) = P_{nt+1}$$

---

[11] Thus, an implicit production function for *j*, a withdrawer of natural resources, can be written

$$Q^{tj} = \min\,[K^j_{nQt}, g^t(K^j_{1Qt}, \ldots K^j_{n-1Qt}, K^j_{n+1Qt}, \ldots K^j_{MQt})],$$

where *Q* is the withdrawn resource If we select "derivatives" of this function to be such that (6) and (7) are satisfied, as we are free to do, this specification is not inconsistent with our general model The cost of *g*ᵗ is the "withdrawal cost" described above, while

$$P_{Qt}Q^{tj} = P_{Qt}K^j_{nQ} = P_t C^{tj}_t.$$

as in (40) above Finally, there is the observation that the spot price of withdrawn natural resources has not substantially changed over time (see Barnett and Morse 1963) Thus we can write $P_{Qt} = P_t$ and $Q_t^l = C_t^l = K_{nQt}^l$ for a withdrawer of a natural resource

With a constant spot price of the withdrawn natural resource, an obvious encouragement to natural resource exploitation (when there is a positive marginal rate of time preference) is a subsidy which is a fixed percentage of the revenue from the sales of the withdrawn natural resource For the producer would rather have a given subsidy this year than next Thus we set

$$X_{nt} = \lambda P_t K_{nt}^{lm} \tag{41}$$

The λ is the "percentage depletion rate"[12] Hence,

$$T_b^{nl} = \sum_t b_{nt}[P_t C_n^l (K_{nt}^l - K_{nnt}^{ll}) - \lambda P_t (K_{nt}^l - K_{nnt}^{ll})] \tag{42}$$

Using (39) and (40),

$$T_b^{nl} = \sum_t b_{nt}(P_{nt+1} - \lambda P_t)(K_{nt}^l - K_{nt+1}^l) \tag{43}$$

Assuming $b_{nt} = b_n$ and using (40) and (19),

$$\begin{aligned} T_b^{nl} = {} & b_n(P_{n2} - \lambda P_1)K_{nl}^{l*} \\ & + \sum_t b_n \left[P_{nt+1}(1 + a_{nt+1}) - \lambda P_t \left(\frac{1}{1 + \rho_t} \right) \right] K_{nt+1}^l \\ & - \sum_t b_n (P_{nt+1} - \lambda P_t) K_{nt+1}^l, \end{aligned} \tag{44}$$

or

$$T_b^{nl} = b_n(P_{n2} - \lambda P_1)K_{nl}^{l*} + b_n \sum_t \left(a_{nt+1}P_{nt+1} + \frac{\lambda \rho_t P_t}{1 + \rho_t} \right) K_{nt+1}^l \tag{45}$$

Hence, for an optimal tax, letting the first lump-sum term in (45) be absorbed in T^{ol} and using (32),

$$a_{t+1}^o = b_n^o \left(a_{t+1}^o + \frac{\lambda^o \rho_t}{(1 + \rho_t)} \frac{P_t}{P_{nt+1}} \right) \tag{46}$$

---

[12] The traditional, Harberger (1955) analysis of such percentage depletion allowances assumes that manufacturing and oil "investments" should be taxed equally if they generate the same streams of cash income It fails to recognize that if the oil "investment" is not undertaken, there is still an accumulation of oil reserves, which is a true social investment Therefore, it is necessary to net the disinvestment of oil reserves out of Harberger's oil "investment" before taxes should be equated on his equal investments The Harberger study should also be corrected for the fact that a depletion allowance is capitalized in the value of the land, thus serving to increase the costs as well as the returns to current oil "investments" Making these adjustments in Harberger's analysis and making intertemporal investment possibilities and taxes explicit leads to our own analysis

Assuming that the income tax rate is that optimal rate applied to equipment, and again assuming that $a^{o}_{t+1} = a^{o}$, we find, substituting (38) into (46), that

$$\lambda^{o} = \frac{P_{nt+1}}{P_t} \tag{47}$$

Hence, the optimal percentage depletion allowance is the ratio of the market value of the conserved natural resource to the value of the product obtained by currently exploiting the resource

The most reliable data on natural resources we have cover the oil industry Here, the ratio of mineral right value to output value has been relatively constant at about 23 percent (see Rooney [1965]) This mineral right value is obtained by adding amortized oil lease payments to royalties The current U S percentage depletion allowance is 22 percent There [illegible] thus a close correspondence in this industry between the optimal and actual percentage depletion allowances[13] While precise data are not available, there is also a correspondence between the low 5 percent depletion allowances given to producers of gravel, peat, pumice, shale, and stone and the obviously low value of mineral rights to these natural resources relative to the prices of the withdrawn resources to consumers And in the extreme case in which mineral rights are essentially free--such as for soil, dirt, moss, minerals from sea water, and air--there is a zero percentage depletion allowance

We now estimate the optimal allowance for the minerals industry in the aggregate Note first that because of the constancy over time of the spot price of withdrawn natural resources, spot withdrawal costs must fall over time so as to make the spot price of a natural resource rise at the productive rate of interest In particular the percentage reduction in spot withdrawal costs over time times withdrawal costs relative to the spot price of a particular natural resource at withdrawal must equal the rate of interest Therefore, the price of a natural resource at withdrawal relative to the withdrawn resource equals the capitalized rate of cost reduction divided by one plus this capitalized rate Since the rate of decrease of withdrawal costs in the minerals industry is about 2 percent per annum (Barnett and Morse 1963), and we are using a productive interest rate of 10 percent, the price of natural resources at withdrawal relative to withdrawn resource price over all valuable minerals is estimated to be 16.67 percent. In fact, the bulk of the statutory depletion allowances fall between 14 and 22 percent (with effective rates slightly lower because of a limitation of the allowance to 50 percent of the net income of the taxpayer).

---

[13] Oil producers are also allowed to write off most of their capital expenses (their "intangible drilling expenses") as current expenditures rather than depreciating them over the life of the producing wells This write-off can be justified by the "national defense" argument mentioned in fn 4, that is, the efficient price controls that are applied during a war imply an insufficient amount of domestic capacity for wartime in the absence of special subsidies Based upon the frequency of past wars and the extent of wartime price controls on oil, a conservative estimate of the optimal subsidy to domestic drilling is 20 percent of the peacetime price of crude oil (see Thompson [1968]) The value of the write-off of intangible drilling expenses can be estimated at around 17 percent of the normal price of crude oil (see Rooney [1965]) The extension of the formal model to admit the possibility of future, successful, defensive wars for a given country is straightforward.

The observed spot price of natural resources relative to withdrawn resources may be constant over time despite its increase for a given natural resource because the quality of the resources exploited may decrease over time. For example, the ratio of oil royalties to the value of the withdrawn oil, a well-known empirical constant, could never have remained constant had not the more easily withdrawable oil deposits been tapped at the earlier dates.

Consumer Durables

Consumer durables, like producer durables, are sold when they are originally created but, unlike producer durables, do not create future benefits for others. There is therefore no "income" from consumer durables to tax. Hence, an excise tax on the production of consumer durables goods is in order. Using again an interest rate of 10 percent, equation (37), and an optimal income tax rate of approximately 25 percent, we have an optimal capital tax rate of 2.5 percent. Therefore, consumer goods lasting 5 years and depreciating in a sum-of-years digits fashion, should be taxed at an initial excise tax rate of

$$2.5\text{ percent}\times\left(1+\frac{10}{15(1.1)}+\frac{6}{15(1.1)^2}+\frac{3}{15(1.1)^3}+\frac{1}{15(1.1)^4}\right) = 5.35\text{ percent},$$

and consumer goods lasting 15 years should be taxed at 10.29 percent using these assumptions. Until very recently, U.S. federal excise taxes on consumer durables ranged between 5 percent and 10 percent with the lower rates generally applying to the relatively short-lived goods.[14]

---

[14] For the various excise tax rates, see Commerce Clearing House (1957, Federal Taxes, Subtitle D). For relative depreciation rates, see Prentice-Hall (1971, par. 15,000). As excise tax rates have recently been volatile, there is no simple method of evaluating the post-1964 excise-tax structure. Some federal excise taxes fall substantially on producer durables, the most notable of which are "business" machines (such as typewriters and computers) and cars and trucks. The business machine case is fairly easy to understand once it is recognized that the sellers of the more expensive machines normally avoid capital taxation by renting their outputs. The less expensive machines are frequently used by consumers, so that an excise tax is in order. The same applies to cars and trucks. However, businessmen should be allowed to expense their purchases of this already-taxed equipment.

An important consumer durable that is not federally taxed in the United States is an individual's home And apartment building depreciation write-offs are so generous that, in view of the ease of transferring these buildings, there are also no substantial federal taxes on these consumer durables But local property taxes seem to compensate for these apparent inefficiencies as effective property tax rates typically are about 2 percent per year, which is close to the federal rate on the other durables treated above My guess is that the cumbersome local property tax and the provision of free education to minors is somehow required by the federal government before a locality can exercise local police power or float tax-exempt bonds In this view, the locality is merely an adjunct of the federal government The chief reason for this suspicion is that it is not plausible that freely competing localities would offer free education to a partially mobile, heterogeneous populace or would use the property tax as a means for financing it Yet the proper education of minors is an activity whose federal subsidization is easy to support—not as a collective good—as a good falling within our model which would be privately underproduced without governmental subsidy because the private decision makers (parents) are not entitled to the increase in future productivity which their decisions (education for their minor children) create [15] We shall employ still another implication of the lack of appropriate parental rewards in the following section While it is fairly obvious that the inefficient parental reward structure serves to rationalize special laws against polygamy, prostitution, divorce, child labor, the minimum wage for teenagers, and social security (which removes a parent's inefficient incentive to instill feelings of guilt in his children in order to illicit their financial support after his retirement), the traditional, inappropriate theory of public finance has buried the important effects of an inappropriate parental reward structure on the efficiency of the tax system in a sea of imagined inefficiencies

Although local property taxes in the United States are also observed to fall on the plants of small businessmen, these capital taxes are roughly offset by the national investment credit And while larger, corporate enterprises are normally able to bargain tax breaks out of small localities in some form or other and therefore effectively avoid local property taxes on their plants, we will find below a substantial reason for applying the current investment credit to U S. corporations

---

[15] Parents may capture part of the increase in productivity that their decisions create by (inefficiently) training their children to feel sufficiently indebted to transfer some future income to their parents, but even then the parent generally has a bias toward educating the child in ways which make it relatively easy to capture returns So privately provided education would generally surrender too much of the child's leisure to "character building," which makes the children feel indebted to the parents and would overly train the child toward occupations with relatively low nonpecuniary rewards.

However, parents who have sufficient utility for their children's utility that they give their children unconditional gifts or inheritances, will make efficient child-rearing decisions (see Ruhter [1974]). Thus, public education need not accommodate the rich who are generally observed to make some unconditional gifts to their children.

Human Capital

"Human capital" as used below is coveted human capital, or "skill," that part of one's human capital stock which he uses to produce goods for the market (or to produce future skill) in value exceeding his subsistence Newly created skills are reproduced in each future period through the worker's taking care of himself and making any necessary expenditures to retain his skill, the latter being treated as the former for tax purposes by granting it a write-off as a current expense

Since skill carries its own maintenance out of what would otherwise not have been coveted capital, its value to the aggressor in each period is the present value of its entire future product Thus, optimal capital taxes on i's skill amount to

$$T^{sio}_{a} = \sum_{a} a^o_t K^i_{st} \sum_{\tau = t} P_\tau C^\tau_s \frac{1}{1 + a^o_\tau} \quad \text{[16]} \tag{47}$$

On the other hand, income taxes are

$$T^{si}_{b} = \sum_{t} b_{st} P_t C^t_s (K^i_{st} - K^{ii}_{sst}) \tag{48}$$

We assume a stationary optimum so that $a^o_\tau = a^o$, $C^\tau_s = C_s$, and $\rho_t = \rho$ We consider an accumulation of a durable skill from the time v onward equal to ΔK^i_{sv}, where $v > 1$

First consider the case in which the accumulation in v is not the result of foregoing income from skill so that $\Delta K^{ii}_{sst} = 0$ for all t Then,

$$\Delta T^{sio}_{a} = \frac{a^o P_{v-1} C_s \, \Delta K^i_{sv} (1 + \rho)}{(1 + a^o)\rho^2} \tag{49}$$

and

$$\Delta T^{si}_{b} = \frac{b_s P_{v-1} C_s \, \Delta K^i_{sv}}{\rho} \tag{50}$$

where $b_{st} = b_s$
Hence, equating ΔT^{sid} and $\Delta T_b{}^{si}{}_3$

$$b^o_s = \frac{a^o (1 + \rho)}{(1 + a^o)\rho} \tag{51}$$

---

[16] In a case in which produced skill is maintained out of that skill itself, eq (47) would, of course, take the standard form of eq (15) above It is easy to verify that our results below also hold to a close approximation for that case (One need only note below that in such a case $K^{ii}_{sst}(1 + \rho_t + a_t) = K^i_{st}$ for $t > v$ so that [50] and [53] are correspondingly reduced to approximately match the reduction in [49])

Tax write-offs for expenditures on education, job search, and worker-owned equipment would substantially subvert this efficient income tax The optimal income tax in (51) is only slightly different from the optimal income tax on purchased producer durables

Now consider the case in which the initial accumulation of a durable skill is accomplished by foregoing income from current skill and keeping leisure time constant Here, (49) remains the same, but

$$\begin{aligned}\Delta T_b^{sl} &= -b_{sv-1}P_{v-1}C_s(\Delta K^l_{ssv-1}) + \sum_{t=v} b_{st}P_tC_s\,\Delta K^l_{sv} \\ &= -b_{sv-1}\frac{P_{v-1}C_s\,\Delta K^l_{sv}}{1+\rho+a} + \sum_{t=v} b_{st}P_tC_s\,\Delta K^l_{sv}\end{aligned} \tag{52}$$

Assuming $b_{st} = b_{sv}$ for $t > v$,

$$\Delta T_b^{sl} = -b_{sv-1}\frac{P_{v-1}C_s\,\Delta K^l_{sv}}{1+\rho+a} + b_{sv}\frac{P_{v-1}C_s\,\Delta K^l_{sv}}{\rho} \tag{53}$$

Using (49), for optimal taxation, we have,

$$\frac{a^o(1+\rho)}{(1+a^o)\rho} = b^o_{sv} - \frac{b^o_{sv-1}\rho}{1+\rho+a^o} \tag{54}$$

Equation (54) states that an optimal income tax is progressive, as the marginal tax rate applying to increases in future wages, b^o_{sv}, must exceed the average tax rate, b_{sv-1}, which one avoids by training rather than working in the $v - 1$st period For individuals whose marginal investments in human capital appear at an average level of foregone earnings, we use (51) and set $b^o_{sv-1} = a^o(1+\rho)/\rho(1+a^o)$ so that (54) can be written as

$$b^o_{sv} = b^o_{sv-1}\left(1 + \frac{\rho}{1+\rho+a}\right) \tag{55}$$

Using 10 percent time-preference rates, year investment periods, and 2 5 percent optimal capital tax rates, we can fit (55) to actual data on U S tax rates[17] The result is that the actual U S marginal income tax rates are within one percentage point of the optimal rates for all approximately average levels of foregone annual income (those between $5,000 and $10,000)

The fit is again very close The actual U S marginal income tax rates are always within 2 percentage points of these theoretically optimal rates for all reasonable levels of human capital investment (i e , for all levels of foregone annual incomes of $25,000 or less) Equation (55) implies higher tax rates on the leisure-produced accumulation of human capital of individuals with above-average foregone earnings on their marginal investments and lower tax rates on individuals with lower than average foregone earnings But it is not implausible that the truly optimal tax rate on leisure-produced skill rises with income rather than remaining at the constant expressed in (51) because the excess of the foregone earnings of an individual over average foregone earnings is a plausible measure of the extent of his parental overtraining to surrender leisure (as the reward structure given parents induces them to instill greater lifetime estimates of the value of work relative to leisure in their young children than they would if they could collect as much of their child's leisure benefits as they can the child's work benefits), while the deficit of an individual's foregone earnings below average foregone earnings is a plausible measure of an individual's parental undertraining concerning the value of work and investing in human capital relative to consuming leisure So we assume, rather crudely, that the optimal tax rates on leisure-produced skill are approximated by the degree of progression implied by the successive application of (55) across all levels of income Then, since the analysis of (49)–(54) with proportionately adjusted optimal tax rates reveals that (55) still describes the optimal taxation of skill accumulation by foregoing current income, (55) describes an optimal degree of progression over all levels of income

---

[17] Future tax rates were computed assuming the individuals will be married, while current rates, b_{v-1}, were computed assuming the investor is single (Lasser 1971)

The above analysis applied to accumulating durable skill so that depreciation could be ignored This appropriately describes most education and training, as most education and training is undertaken by young people who are going to use their training for very many years However, some training is undertaken by older people and some training of youth is for short-term or risky careers In these cases, the above tax on human capital accumulation is too high Some sort of subsidy for retraining or for training in occupations with short durations is called for A depreciation allowance on human capital investments would be insufficient For since the wage increases required to justify a short-term human capital investment are greater than those required to justify a durable investment, the progressive income tax effects a greater tax rate on short-term human capital investments than on long-term investments We do observe a tax break for risky and short-term human capital investments in the form of an "income averaging" opportunity, but we have made no attempt to quantify the effects of this tax provision

Producer Durables Used by the Original Producer

An original producer of producer goods may avoid the taxation of his produced capital by writing off its cost of original production as a current expense and then using it himself for future production. The difficult problem of the government is to detect these activities and to apply special taxes whenever such activities occur. We now let the progressive income tax extend over from human to nonhuman capital but also apply only half of the ordinary income tax rate to "capital gains" income, income from the sales of the outputs of producer goods that can be used by the original producers of the goods The most wealthy individuals are thereby drawn into industries in which produced producer goods can be used by the original producer, thus driving the less wealthy out of these activities The wealthy are induced by the tax break to identify the capital gains situations, those where they have produced their own producer goods, and the government may now treat the previous expenses of producing this capital as capital expenditures, still leaving the wealthy better off than with ordinary-income treatment of both expenditures and receipts and with a higher value on the entire enterprise than the less wealthy individuals [18] The resulting tax on producer goods used by the original producer is therefore about half as large as the highest tax bracket, reduced by about 25 percent, when, as in the United States, the capital gains tax is collected at the time of sale rather than as the gains accrue The highest tax bracket in the United States is currently 70 percent, which puts the effective U S tax rate on the income from producer goods used by the original producer at about its optimal level of 25 percent

However, the above procedure efficiently taxes only the producers' *originally* produced capital that he uses for himself It does not tax the internal accumulation of the successive capital outputs of the originally produced capital We have already seen in the case of human capital that in order to tax capital accumulated by foregoing current income, some sort of progressive tax schedule is required In the present case the relevant tax brackets are very high, so that significant progression is no longer possible One solution to this problem, which is informationally impractical, is to tax capital gains as they accrue, at about one-third of the ordinary income tax rate More practically, we may tax realized capital gains at tax rates which increase to the date of realization with the real rate of interest, starting from one-third of the ordinary income tax rate However, this latter solution would remove the present discouragement to *transact* in the type of capital covered by capital gains taxation (This discouragement is also reflected in the harsh tax treatment of capital losses and realized short-term capital gains) A rational basis for this discouragement, although not contained in the current model, is that there is a redistributional element in most transactions involving capital gains, one in which the buyer is more optimistic than the seller regarding the future value of the goods The tax discouragement therefore serves to save the socially wasteful transaction costs involved in these transactions. Thus, we have no practical way clearly to improve the capital taxation of the capital output of those producer goods which an individual produces for his own use.

---

[18] When the government fails to treat the previous expenses of the investor as capital expenditures, an inefficient "tax haven" exists—such as has recently been the case in the United States in cattle, timber, and crop development. The U S government ordinarily shuts down these tax havens, although rather slowly

There are numerous cases in which capital gains treatment cannot be afforded to the sale of the goods produced by capital that the producer produces for himself. These occur when there is a continual stream of marketed outputs and expenditures over time, so that it becomes difficult, if not impossible, to attribute any measure of marketed outputs to the inputs that produce them. Thus, for example, a producer's maintenance expenditures are treated not as capital costs but as current business expenses, while the returns from such expenditures are treated not as capital gains but as ordinary income when the final products are marketed. In order to avoid the application of the high progressive tax rates to the nonhuman capital of individuals with continual streams of marketed outputs and expenditures on such capital, we allow these individuals to incorporate their nonhuman capital, taking their returns from the corporation in a form in which they pay only about a 10 percent tax rate.[19] Our problem is then to specify an efficient tax on corporations in light of their ability to produce their own capital. The simplest policy is to tax the value of the company (or the value of its stock, still taxing the interest income of the creditors) at the efficient, 2–3 percent capital tax rate and drop the tax on dividends and capital gains. Such a policy would directly tax any capital in the company—regardless of how it is accumulated. The costly implementation of realistic depreciation and efficient depletion allowances would be avoided, as would the taxation of dividends, capital gains, and corporate profits. But we assume that such a policy is not available, or, if it is, that corporations remain whose stock does not trade at observed prices. The problem then is to specify an efficient corporation income tax in view of the corporation's ability to produce some of its own capital. Now purchased capital is recorded on the books of a company as an asset, but internally produced capital which the company expenses is not included on the books as an asset. However, this latter kind of capital is included in the market evaluation of the company so that the depreciated stock of externally purchased assets in a corporation relative to its total capital stock can be represented by the ratio of the company's book value to its market value. We assume this is constant within each corporation.

Letting B_{ft}/K_{ft} represent the ratio of book to market value of the assets of the firm, replacing K^{i}_{et} in (35) with B^{i}_{ft}, using (32) and (36), and adjusting for a 10 percent effective tax on dividends and capital gains, the optimal corporation profits tax is

---

[19] Partial dividend exclusions and taxes on realized capital gains at about half the tax rate applicable to ordinary income seem to achieve about this effective rate for the typical investor. The number may seem a little low, but it reflects the significant advantages to delaying realized gains, and giving charity and bequests in the form of appreciated stock.

$$b_f^\circ = \left(\frac{K_f}{B_f}\right)\frac{b_e^\circ}{1.10} \tag{57}$$

The average K_f/B_f has been estimated to be 1.6.[20] Thus, given our other estimates, the average optimal tax rate on corporate income is about 37 percent. While this is somewhat lower than the statutory rate of 48 percent for a large company, the presence of the 7 percent investment tax credit in recessions has served to substantially lower the effective tax on produced producer durables. Assuming that half of the years are "recession" years, the annual investment credit 3½ percent. And assuming that the typical age of produced producer durables is 10 years, the optimal effective excise tax (using the formula for consumer durables) on corporate purchases of producer durables is about 7 percent × 1.6. This means that the effective 3½ percent investment credit subsidizes capital by about 30 percent of the optimal capital tax. But since the corporate profits tax rate is greater than the optimal rate by about 30 percent of the optimal rate, the combined 48 percent corporate profits tax and 7 percent investment credit in recessions effects very close to an optimal tax on producer durables in corporations.

Surrenderable Capital

Apparently, the only empirical cases in which the costs of defending certain units of capital are significantly higher than the costs of defending equally valuable coveted capital are cases in which the capital is located on foreign soil. In accord with our general results, the U.S. government neither substantially commits itself to defend such capital nor significantly taxes the income from such capital.[21] Rather, the United States sells insurance against foreign confiscation to the various companies with foreign holdings, in effect charging a price for providing only those protection services that it would rationally supply as an insurer in order to avoid claims by the insured much as private medical insurers rationally supply proper medical care in order to avoid future insurance claims

---

[20] This was done by multiplying an estimate of the rate of return to book value of equity for U.S. manufacturing corporations in 1966 (Pechman 1971, p. 307) by an estimate of the price-earnings ratio for U.S. industrials for the same year (Moody's Investors' Service 1967, p. a 23). This was then adjusted to represent the ratio of the book value of companies relative to their market value by adding on the ratio of debt to net worth and dividing by one plus this ratio (Pechman 1971, p. 307).

[21] An exception applies to overseas earnings of U.S. companies when taxes paid to foreign countries are below what domestic taxes would be. The United States taxes such earnings as domestic earnings and credits the foreign taxes against the tax bill. This is a significant exception only when foreign tax rates are very low, in which case the United States is probably protecting the foreign country, so that the domestic U.S. tax rates should apply.

III. Conclusion

Our results indicate that the U.S tax structure does not produce the deluge of malincentives and economic inefficiencies that one finds when using the traditional theory of public finance Rather, using a more accurate theory of the nature of national defense, our tax structure produces roughly optimal incentives Because no model such as our was available to the advisors or decision makers during the development of the U S tax structure, our results strongly suggest that our political system, using the self-interested calculations of its citizens, has somehow been able to systematically produce a substantially more efficient tax structure than our economists, using the traditional theory of taxation, have been able to recommend

While incentives under the U S tax system appear to be remarkably efficient, we are not advocating the status quo In particular, we have argued that the current U S tax system could be improved by reinstating excise taxes on those consumer durables which are not used for business purposes and replacing corporate income taxes on companies with publicly traded shares with an annual tax of about 2½ percent on the market value of their common and preferred stock, dropping the personal income tax on dividends and capital gains on such stock

References

Barnett, H J , and Morse, C *Scarcity and Growth* Baltimore Johns Hopkins Univ Press, 1963

Commerce Clearing House *Code of Regulations* Chicago Commerce Clearing House, 1957

Harberger, A "The Taxation of Mineral Industries' *Federal Tax Policy for Growth and Stability* Washington Joint Econ Committee of Congress, Government Printing Office, 1955

Knight, F H "Professor Hayek and the Theory of Investment " *Econ J* 45 (December 1935) 77–94

Lasser, J K *Your Income Tax* New York Simon & Schuster, 1971

Malinvaud, E "Capital Accumulation and Efficient Resource Allocation." *Econometrica* 21 (April 1953) 233–68

Marshall, A. *Principles of Economics* 8th ed New York Macmillan, 1925

Moody's Investor Service. *Moody's Industrial Manual* (June 1967)

Pechman, J *Federal Tax Policy* Washington, D C Brookings Inst., 1971

Prentice-Hall *Federal Taxes* Vol. 2 Englewood Cliffs, N J Prentice-Hall, 1971

Rooney, R. F. *Taxation and Regulation of the Domestic Crude Oil Industry* Ph D dissertation, Stanford University, 1965

Ruhter, W "Childhood, Human Capital Development, and Parental Incentives." Unpublished manuscript, UCLA, 1974

Samuelson, P A "The Pure Theory of Public Expenditure" *Rev. Econ. and Statis* 36 (November 1954) 387–89

——— "Tax Deductibility of Economic Depreciation to Insure Invariant Valuations" *J P E* 72 (December 1964) 604–6

Smith, A. *Wealth of Nations* Vol. 2 New York Random House, 1937.

Thompson, E A "Statement" *Governmental Intervention in the Market Mechanism The Petroleum Industry* Hearings before the Subcommittee on Antitrust and Monopoly of the Committee on the Judiciary, May 1968

Vertical Control with Variable Proportions

Frederick R Warren-Boulton

Washington University, Saint Louis

If a monopolized input can be used in variable proportions, vertical integration or an equivalent tying arrangement will increase monopoly profits Assuming a constant elasticity of final demand and a constant-elasticity-of-substitution production function for the final product, this paper uses a combination of analytic and simulation techniques to show that vertical control will also reduce the demand for nonmonopolized inputs and, except under particular cost conditions in the input industries, will result in an increase in the price of the final product The direction of the combined welfare effect, however, depends on the specific parameter values

If a monopolized factor or intermediate good can be used in variable proportions with other inputs by downstream producers, pricing of the monopolized input above its marginal cost will distort the input ratio chosen by downstream users Vertical control[1] by the input monopolist can prevent this distortion and thus increase monopoly profits This proposition has been established several times in the literature, notably by McKenzie (1951) and by Burstein (1960), and a graphical formulation has been presented by Vernon and Graham (1971)

For public policy, however, the critical problem is to determine the effects of vertical control on the prices of the inputs and the final product

---

I would like to express my particular appreciation to Charles Berry for his encouragement and assistance I would also like to thank Trout Rader, Peter Sturm, and a referee for their helpful comments

[1] Vertical control can be exerted by firms, unions, or individuals in the form of vertical integration (complete control through ownership) or in more partial forms such as tying or royalty arrangements The use of a monopolized input in variable proportions is one of several possible causes of vertical control Other reasons include economies of control and information, reduction of risk, price discrimination, avoidance of regulation or price controls, elimination of bilateral or successive monopoly, and the raising of entry barriers For a comprehensive discussion of possible reasons for vertical control, see Williamson (1971)

and hence the effect on welfare[2] This paper examines these effects, assuming a constant-elasticity-of-substitution (CES) downstream production function and a constant elasticity of demand for the final product A combination of analytic results and simulation experiments is used to show that, except under particular cost conditions in the input industries, vertical control results in higher prices for the consumers of the final product and reduced demand for nonmonopolized inputs The direction of the "combined" welfare effect on consumers, producers of nonmonopolized inputs, and monopoly profit receivers, however, depends on the specific parameter values in the demand and production functions

The paper is divided into three sections Section I assumes constant marginal costs in the input industries Section II extends the results to cases of nonconstant input marginal costs Section III examines the effect of relaxing other assumptions and attempts to draw some conclusions for public policy

I. Vertical Control with Constant Input Costs

Assume that a final product, X, is produced by competitive firms using two inputs The first input, A, is monopolized, the other input, B, is supplied competitively The production function for the final product is of constant elasticity of substitution, σ, and constant returns to scale

$$X = Y[\sigma A^{(\sigma-1)/\sigma} + (1-\sigma)B^{(\sigma-1)/\sigma}]^{\sigma/(\sigma-1)},$$

$$Y > 0, 0 < \delta < 1, 0 \leq \sigma \leq \infty \quad (1)$$

Assume also a constant-elasticity demand function for the final product of the form

$$X = Z/P_x^{\eta}, \qquad Z > 0, \quad \eta > 1 \qquad (2)$$

Throughout this section, it will also be assumed that the marginal cost of producing the monopolized input, M_a, is constant over the relevant range, and that B is available in infinitely elastic supply at price P_b

The first step in determining the effect of vertical control on the price of the final product is to derive the profit-maximizing price for A before vertical control Since the supply of A is monopolized,

---

[2] While Vernon and Graham assume that the price of the final product is unchanged after integration, Burstein's opinion appears to be that the price of the final product will rise after integration or an equivalent tying arrangement, since "the very existence of substitution possibilities will in general prevent the monopolist from facing the competitive producers of the final product with production costs high enough to *force* them to charge a price (for the final product) equal to the monopoly price" (Burstein 1960, p 81) In a recent article in this *Journal*, Schmalensee (1973) examines the effects of vertical integration with variable proportions under assumptions of a linear homogeneous production function for the final product and constant costs for all inputs The implication of Schmalensee's results is that the price effect of vertical control cannot be signed without additional information on the cost and demand conditions

$$P_a = \frac{M_a}{1 - 1/E}, \tag{3}$$

where E, the elasticity of derived demand for A, is given by equation (4) [3]

$$E = \frac{\sigma(\eta + e) + k_a e(\eta - \sigma)}{\eta + e - k_a(\eta - \sigma)} \tag{4}$$

where e is the elasticity of supply of B, and k_a is the share of the cost of A in the total cost of producing X Since B is assumed to be available in infinitely elastic supply, $e = \infty$, and equation (4) reduces to

$$\underset{e=\infty}{E} = k_a\eta + (1 - k_a)\sigma \tag{5}$$

From the first-order conditions for cost minimization,

$$k_a = \frac{1}{1 + [(1 - \delta)/\delta]^{\sigma}(P_a/P_b)^{\sigma - 1}} \tag{6}$$

Substituting equations (5) and (6) into equation (3),

$$(\eta - 1)P_a + \left[\left(\frac{1 - \delta}{\delta}\right)^{\sigma}\left(\frac{\sigma - 1}{P_b^{\sigma - 1}}\right)\right] \times P_a^{\sigma} - \left[\left(\frac{1 - \delta}{\delta}\right)^{\sigma}\left(\frac{\sigma}{P_b^{\sigma - 1}}\right)\right]P_a^{\sigma - 1} - \eta M_a = 0 \tag{7}$$

Equation (7) cannot readily be solved explicitly for P_a We do not, therefore, have an expression for P_a which can be used in further derivations However, P_a can be computed over the range of the various parameters, notably η, δ, and σ Since the downstream X industry is assumed competitive, the price of X is equal to its marginal cost Thus, once P_a is determined, we can solve for P_x [4]

$$P_x = \frac{1}{Y}\left[P_b^{1-\sigma}(1 - \delta)^{\sigma} + P_a^{1-\sigma}\delta^{\sigma}\right]^{1/(1-\sigma)} \tag{8}$$

After vertical control, an unconstrained integrated monopolist will set marginal revenue equal to marginal cost in the X market The resulting price for X will be

$$P_x = \frac{\bar{M}_x}{1 - 1/\eta} = \left[\frac{P_b^{(1-\sigma)}(1 - \delta)^{\sigma} + M_a^{(1-\sigma)}\delta^{\sigma}}{Y(1 - 1/\eta)}\right]^{1/(1-\sigma)} \tag{9}$$

---

[3] See Hicks (1963, p 244)

[4] For a derivation, see Nerlove (1967, p 111)

The A monopolist may not, however, be able to set a price of $\hat{P}_x$ for the final product after integration The integrated firm cannot set a price for X which is higher than the independent downstream firms' cost of producing X The cost of producing X for an independent producer, a function of the price of A set by the A monopolist, is thus the "entry price of X," $\hat{P}_x$ If the A monopolist is prepared to raise the price of A after vertical control to any arbitrarily high level, $\hat{P}_x$ becomes the cost of manufacturing X using only B (or, equivalently, the value of P_x in eq [8] with $P_a = \infty$) If $0 \leq \sigma \leq 1$, $\hat{P}_x$ is infinite, since X cannot be produced without using A But for $\sigma > 1$, $\hat{P}_x$ is less than infinity,[5] and may even be less than $\hat{P}_x$ Thus the ratio of the price of the final product after vertical control to the price of the final product before vertical control is

$$R = \frac{\hat{P}_x}{P_x} \leq \frac{\hat{P}_x}{P_x} \qquad (10)$$

R can easily be shown to be equal to unity if the elasticity of substitution is zero or infinity, since, for either of these two extreme values for σ, factor proportions cannot be affected by vertical control [6] In addition, for the special case of $\sigma = 1$, it is possible to derive an explicit expression for R[7] which, with constant costs in both input industries, reduces to

$$\underset{\substack{\sigma = 1 \\ e = g = \infty}}{R} = \frac{(1 - 1/E)^{\delta}}{(1 - 1/\eta)} > 1, \qquad (11)$$

---

[5] For the CES production function with $\sigma > 1$, the average product of B approaches a positive lower limit as the A/B ratio approaches zero (see Arrow et al [1961, pp 230–31]) Hence production is possible with B alone, and the unit cost of producing X using only the B input, $\hat{P}_x$, is a finite upper limit to the integrated monopoly's price

[6] See Burstein (1960) or Vernon and Graham (1971)

[7] Assuming a Cobb-Douglas production function for the final product $X = YA^{\delta}B^{(1-\delta)}$, $Y > 0$, constant elasticity of demand for the final product $X = Z/P_x^{\eta}$, $Z > 0$, $\eta > 1$, constant elasticity of supply for the cooperant input $B = CP_b^{e}$, $C > 0$, and a constant-elasticity marginal-cost schedule for the monopolized input $A = DM_a^{g}$, $D > 0$, we can solve for the price of the final product before vertical control

$$P_x = \left\{ \frac{Z^{\delta(e-g)+e+1}}{Y^{(e+1)(g+1)}C^{(1-\delta)(g+1)}D^{\delta(e+1)}(1-\delta)^{e(1-\delta)(g+1)} \times [\delta(1 - 1/E)]^{\delta g(1+e)}} \right\}^{1/[\eta(\delta e+g+1-\delta g)+g(\delta+e)+e(1-\delta)]}$$

and the price of the final product after vertical control

$$\hat{P}_x = \left[\frac{Z^{\delta(e-g)+e+1}(1 + 1/e)^{e(1-\delta)(g+1)}}{Y^{(e+1)(g+1)}C^{(1-\delta)(g+1)}D^{\delta(e+1)}(1-\delta)^{e(1-\delta)(g+1)} \times \delta^{\delta g(e+1)}(1 - 1/\eta)^{g(e+\delta)+e(1-\delta)}} \right]^{1/[\eta(\delta e+g+1-\delta g)+g(\delta+e)+e(1-\delta)]}$$

The price ratio is thus.

$$\underset{\sigma = 1}{R} = \left[\frac{(1 - 1/E)^{\delta g(1+e)}(1 + 1/e)^{e(1-\delta)(g+1)}}{(1 - 1/\eta)^{g(e+\delta)+e(1-\delta)}} \right]^{1/[\eta(\delta e+g+1-\delta g)+g(\delta+e)+e(1-\delta)]}$$

where

$$\underset{\substack{\sigma=1\\ \epsilon=\infty}}{E} = 1 + \delta(\eta - 1)$$

and g = elasticity of the marginal-cost schedule in the A industry (assumed infinite in Section I) For this special case, R is always greater than unity and approaches unity as η approaches infinity or as δ approaches one Thus the price of the final product must go up after integration if the elasticity of substitution is unity and there are constant costs in both input industries The price of the final product will go up by a larger proportion after vertical control, the smaller are δ and η Thus the smaller the share of the cost of the final product which is due to the monopolized input, and the less elastic the demand for the final product, the greater the percentage increase in the price of the final product after integration [8]

We can also show that $R > 1$ for a second special case, that of $\sigma \geq \eta$ If $\sigma \geq \eta$ then $E \geq \eta$ [9] The monopoly markup on A before vertical control, P_a/M_a, is thus less than or equal to the markup on X after vertical control, $\hat{P}_x/\hat{M}_x$ Furthermore, an increase in the price of one input by any proportion must result in a rise in the price of the final product by a lesser proportion if the prices of other inputs are held constant Thus P_a/M_a must be greater than the ratio of P_x to the price of X which would exist if there were competition in both input and final-product industries With constant input costs, this competitive price for X, P'_x, is also the integrated monopolist's marginal cost of producing X, $\hat{M}_x$ Thus if $\sigma \geq \eta$,

$$\frac{P_x}{P'_x} < \frac{P_a}{M_a} = \frac{1}{1 - 1/E} \leq \frac{1}{1 - 1/\eta} = \frac{\hat{P}_x}{\hat{M}_x},$$

and since $\hat{M}_x = P'_x$, $P_x < \hat{P}_x$

The conclusion that vertical control results in a rise in the price of the final product for two special cases, $\sigma = 1$ and $\sigma \geq \eta$, leads one to expect this result for the general case of $0 < \sigma < \infty$ The inability to derive an explicit expression for P_a from equation (7) prevents the derivation of an analytic proof The effects of vertical control for the general case of $0 < \sigma < \infty$ can be examined, however, through the use of computer simulations, using a range of values for each important parameter In particular, the effect of vertical control on the price of the final product and on the levels of consumers' surplus[10] and monopoly profits was computed for all combinations of: $\sigma = 0.00, 0.25, 0.50, \quad 4.00$, $\eta = 1.1, 1.5, 2, 3, 4, 5$, $\delta = 0.05, 0.10, 0.25, 0.38, 0.50, 0.63, 0.75, 0.95$ Other parameter values were $Y = 1.00$, $M_a = 1.00$, $P_b = 1.00$, $Z = 1{,}000$ The conclusions from these simulations are as follows.

---

[8] These results for the $\sigma = 1$, $\epsilon = g = \infty$, case are also derived in Schmalensee (1973, pp 447–48)

[9] From eq (4), if $\sigma = \eta$, then $E = \sigma = \eta$ If $\sigma > \eta$, then $E > \eta$, and if $\sigma < \eta$, then $E < \eta$ The elasticity of derived demand is thus a weighted average of σ and η

A The Effect of Vertical Control on the Price of the Final Product

1 With constant costs in both input industries, the price of the final product always increases for elasticities of substitution between zero and infinity

$$\underset{\substack{0<\sigma<\infty \\ \theta=e=\infty}}{R} > 1$$

As the elasticity of substitution is increased from zero, R rises above unity, reaches a maximum, and then declines, approaching unity again as the elasticity of substitution approaches infinity

2 The lower the elasticity of final demand, the greater is $\hat{P}_x/P_x$, although $\hat{P}_x$ and P_x are both decreasing functions of η Thus if the entry constraint is not effective, a lower elasticity of final demand will increase R

The entry price, $\hat{P}_x$, is not a function of η A lower η will, however, increase P_x Thus $\hat{P}_x/P_x$ is an increasing function of η, and, if the entry constraint is effective, a lower elasticity of final demand will decrease R

The effect of varying η is illustrated in figures 1*a* and 1*b*, which plot R as a function of σ for $\eta = 2$ and $\eta = 4$ Figure 1*a* assumes $\delta = \frac{1}{4}$, so that, after some value for σ, the $\hat{P}_x$ constraint is effective Thus, for example, if $\sigma = 1$, an increase in η reduces R, while if $\sigma = 2$, an increase in η increases R In figure 1*b* $\delta = \frac{3}{4}$, and the $\hat{P}_x$ constraint is never effective Thus an increase in η will lower R for all values of σ

3 The effect of varying δ parallels that of varying η A higher value for δ shifts $\hat{P}_x/P_x$ down, but shifts $\hat{P}_x/P_x$ upward Figure 1*c* shows R for $\eta = 2$ and $\delta = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$

These results for η and δ are consistent with the analytic results from the $\sigma = 1$ case (since for $\sigma = 1$, the $\hat{P}_x$ constraint is never effective) and are intuitively plausible *Ceteris paribus*, the less elastic the final demand, and the less "important" the monopolized input (the lower is δ), the greater the difference which can be expected between an unconstrained monopoly over one input and an unconstrained monopoly over a final product On the other hand, if δ is large there is less likelihood of the input monopolist being confronted after vertical integration with

---

[10] It should be noted that the use of consumers' surplus as a measure of welfare requires particular assumptions, for example, homotheticity of preferences (see Rader, 1972 pp. 234–43) For possible ambiguities in the use of producers' surplus later in this paper, see Mishan (1968).

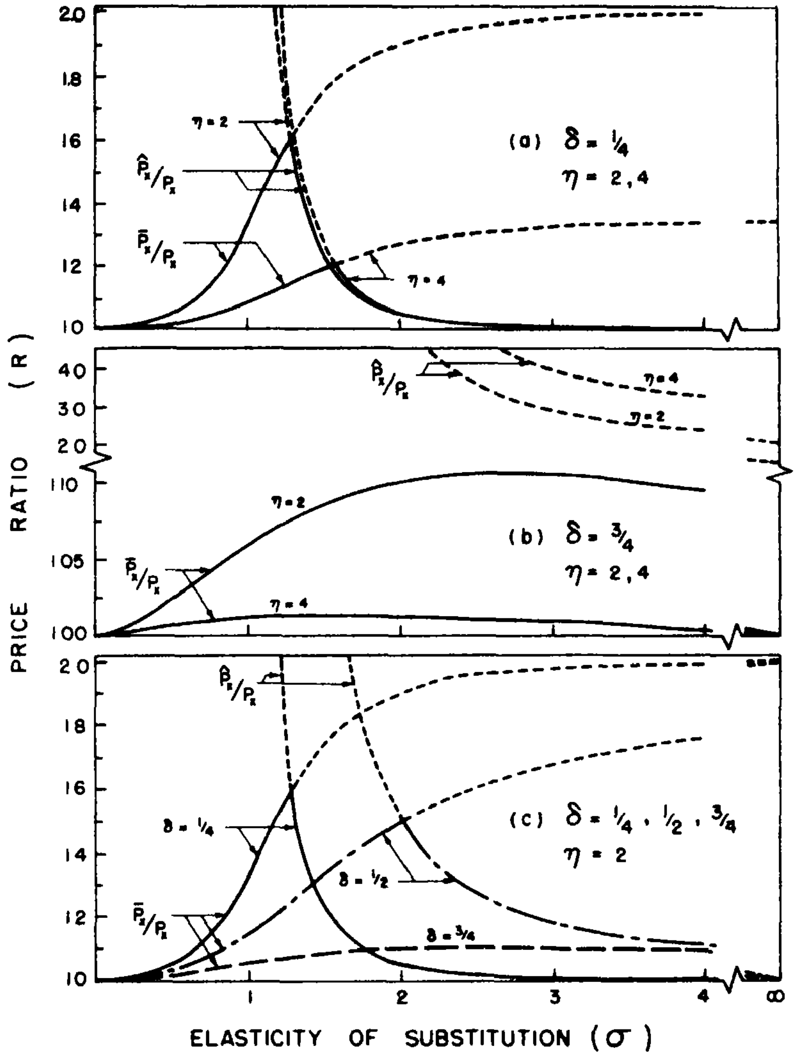


FIG 1 —Effect of vertical control on the price of the final product

effective competition from other firms in the final-product industry (a low $\bar{P}_x$) Similarly, the more elastic the final demand, the less likely that the desired monopoly price for the final product will be high enough to exceed $\bar{P}_x$

Thus the effect of η and δ on R cannot be determined in general without information as to the effectiveness of the entry constraint For $0 \leq \sigma \leq 1$, if the A monopolist is willing and able to raise the open market price of A to any height (ultimately, by simply refusing to sell A at any price), then the $\hat{P}_x$ constraint is not effective and R is a decreasing function of η and δ. Even over this range of σ, however, if the A monopolist does not raise the price of A sufficiently after vertical control, he may be confronted with a maximum price for the final product which is less than the desired monopoly price for the final product.

B. The Effect of Vertical Control on Monopoly Profits and the Incentive for Vertical Control

1. Monopoly profits must increase after vertical control for $0 < \sigma < \infty$. This increase reaches a maximum at some $\sigma > 1$. The lower are δ and η, the greater the magnitude of the increase in profits, and the lower the value for σ where the increase in profits is at a maximum.

While the absolute magnitude of the increase in monopoly profits is relevant for calculating the total welfare effect, a better proxy for predicting the likelihood of vertical control might be the proportion by which monopoly profits could be increased by vertical control, or the increase in profits as a proportion either of the cost of manufacturing A or the total costs of manufacturing X. These three alternative indicators of the likelihood of vertical control are also higher, the lower are δ or η, except that, where the $\hat{P}_x$ constraint is effective, a higher value for δ increases the gain in monopoly profits as a proportion of the costs of A or X.

2. Since an increase in the price of the final product after vertical control may be impossible or undesirable in some cases (because of regulation, threat of antitrust proceedings, possible entry at the A level, or loss of sales of A in other markets if P_a is increased), it is useful to separate the total increase in monopoly profits into two components: the increase in profits which would be possible without changing the price of the final product after vertical control ($EFCY$), and the further increase in profits due to the change in the price of the final product ($\Delta\pi\ \Delta P_x$). If there are constant costs in the B industry, the first component can be regarded as a pure "efficiency" effect. It reflects the increase in efficiency of production of the final product (the reduction in the cost, pricing inputs at their social marginal cost, of producing any given amount of the final product) due to the elimination of the input-ratio distortion.[11]

Graphically, vertical control results in both a downward shift in the cost curve for X and an increase in the price of X (fig. 2). Note that

---

[11] It is this "efficiency" effect which is shown graphically by Vernon and Graham (1971).

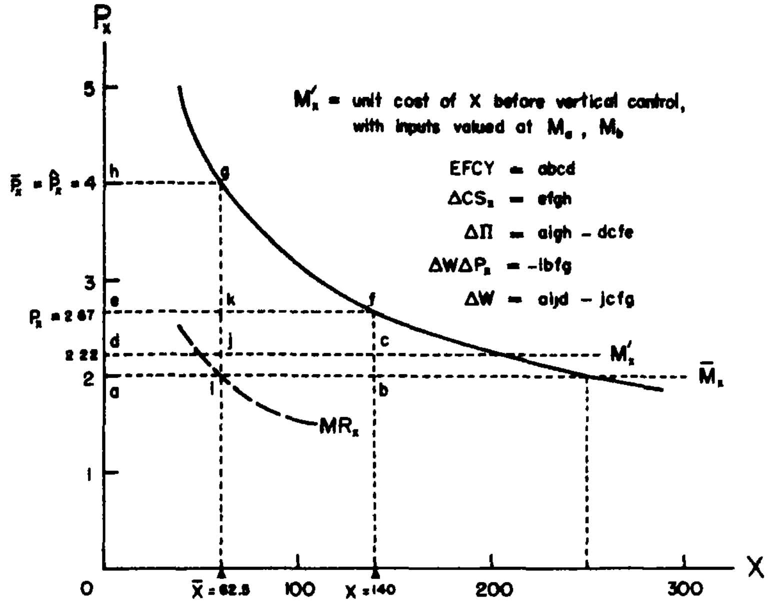


FIG 2 —Effect of vertical control on final-product industry (numerical results assume $\sigma = \eta = 2, \delta = \frac{1}{2}$

because of the previous existence of monopoly power in the input industry, the welfare loss due to the price increase, $\Delta W\ \Delta P_x$, is not just the "welfare triangle," *kfg* in figure 2, but the area *ibfg*

Since the welfare effect of the efficiency component is unambiguously positive, and since it may be the only effect of vertical control, the efficiency component is worth separate examination As a function of σ, δ, and η, the efficiency component shows the same pattern as the total increase in profits As a proportion of the total increase in profits, however, the efficiency gain is higher, the lower is σ and the higher are δ and η As σ rises, and as δ and η fall, the incentive for vertical control increasingly becomes the achievement of market power rather than the elimination of inefficiency

If the price of the final product is prevented from increasing, the likelihood of vertical control occurring, as measured by the ratio of the efficiency gain to monopoly profits before vertical control, or to the costs of *A*, is higher, the lower are δ and η, and the maximum percentage gain is reached at lower σ for lower δ and η As a proportion of the total costs of *X*, the efficiency gain is greater for lower values for η, but, over high values for σ, the proportion is higher for larger values for δ

Whether or not the price of the final product is permitted to rise, therefore, the maximum incentive for vertical control will exist under a combination of very small values for δ and η and an elasticity of substitution only slightly greater than unity As δ and η increase, so does the "optimal" elasticity of substitution

C The Effect of Vertical Control on Welfare

With constant costs in the B industry, the effect on welfare is the combination of the change in monopoly profits and the change in consumer surplus The change in consumer surplus is always negative, while the change in profits is always positive If income distribution effects are ignored, the change in consumers' surplus, ΔCS_x, and the change in profits, $\Delta\pi$, can be lumped together to arrive at a "total" welfare effect from vertical control

$$\Delta W' = \Delta CS_x + \Delta\pi$$

Alternatively, the total change in welfare, ΔW, can be viewed as occurring in two steps (see fig 2 for an illustration) First, holding the quantity of the final good constant, vertical control results in a reduction in the real cost of producing that quantity of the final product This efficiency effect, $EFCY$, is one component of the increase in profits resulting from vertical control In the second step, the integrating monopolist raises the price of the final product, which results in a further increase in monopoly profits, $\Delta\pi\,\Delta P_x$, and a reduction in consumers surplus ΔCS_x The combined welfare effect of this price increase, $\Delta W\,\Delta P_x$ is necessarily negative $\Delta W = EFCY + \Delta W\,\Delta P_x$, $\Delta\pi = EFCY + \Delta\pi\,\Delta P_x$, $\Delta W\,\Delta P_x = \Delta CS_x + \Delta\pi\,\Delta P_x < 0$

The following conclusions can be drawn from simulations of the welfare effects of vertical control

1 As a function of the elasticity of substitution, the effect of vertical control on total welfare, ΔW, rises from zero as σ increases from zero, reaches a positive maximum, declines to a negative minimum, and then rises again For parameter values under which the $\hat{P}_x$ constraint is never effective (as in fig 3a), ΔW approaches zero again from below as $\sigma \to \infty$ If the $\hat{P}_x$ constraint eventually becomes effective at some value for σ (fig 3b), ΔW reaches a second positive maximum before declining again to zero from above as $\sigma \to \infty$

2 If the $\hat{P}_x$ constraint is not effective, lowering η or δ increases the absolute value of the net welfare effect In addition to this scale effect, the value for σ at which the net welfare effect first becomes negative,

$$\sigma_{\substack{\Delta W = 0\\ \partial[\Delta W]/\partial\sigma < 0}} ,$$

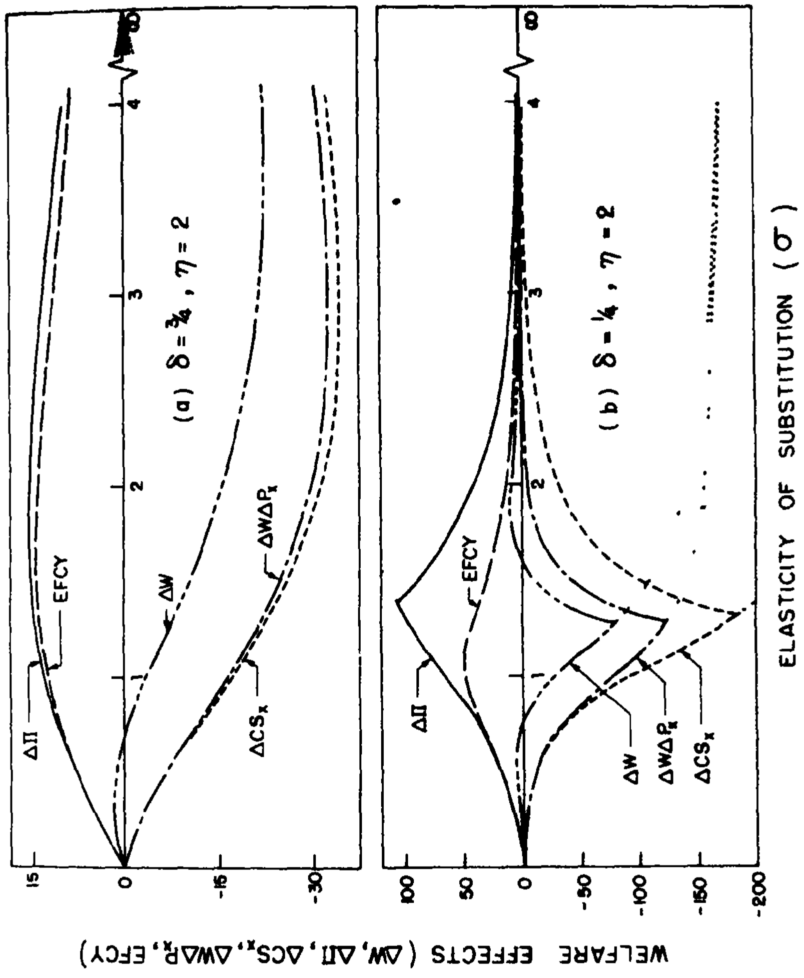


FIG 3 —Effect of vertical control on profits, consumer surplus, and welfare

is higher, the lower are δ or η As $\delta \to 1$, or as $\eta \to \infty$,

$$\sigma_{\substack{\Delta W = 0 \\ \partial[\Delta W]/\partial\sigma < 0}} \to 0$$

As $\delta \to 0$, or as $\eta \to 1$,

$$\sigma_{\substack{\Delta W = 0 \\ \partial[\Delta W]/\partial\sigma < 0}} \to 1$$

Thus if the $\hat{P}_x$ constraint is not effective, the net welfare effect will be negative over a minimum range for σ of $\sigma \geq 1$ Over this range for σ, the absolute value of the welfare effect will be greater, the lower are δ and η For the $0 < \sigma < 1$ range, a lower δ or η will increase the range for σ over which ΔW is positive, and increase the magnitude of ΔW

If the $\hat{P}_x$ constraint is effective, however, the net welfare effect will become positive again at some σ, and the value for σ where ΔW again becomes positive will be lower, the lower are δ or η As $\delta \to 0$ or as $\eta \to 1$,

$$\sigma_{\substack{\Delta W = 0 \\ \partial[\Delta W]/\partial\sigma > 0}} \to 1$$

Thus the effect of lower values for δ and η, if the $\hat{P}_x$ constraint is effective for some σ, is to narrow the range around $\sigma = 1$ where $\Delta W < 0$[12] and increase the scale of the welfare effect Figure 4*a* graphs ΔW as a function of σ for $\delta = \frac{1}{4}$, $\eta = 3, 4$ Figure 4*b* shows $\Delta W(\sigma)$ for $\delta = \frac{3}{4}$, $\eta = 3, 4$ Figure 4*c* shows $\Delta W(\sigma)$ for $\eta = 2$, $\delta = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$

It is thus difficult to draw any general conclusions for the direction of the welfare effect If the particular case is such that it is clear that the input monopolist exerting vertical control would have the power to

---

[12] The net welfare effect is always negative at $\sigma = 1$ if there are constant input costs With $\sigma = 1$, $e = g = \infty$

$$\Delta W_{e = \alpha} = \Delta CS_x + \Delta\pi,$$

$$\begin{aligned}
\Delta W_{\substack{\sigma = 1 \\ e = g = \infty}} &= \frac{Z}{(\eta - 1)} \left[\frac{(2\eta - 1)}{\eta \hat{P}_x^{(\eta - 1)}} - \frac{(2E - 1)}{E P_x^{(\eta - 1)}} \right] \\
&= \frac{Z}{(\eta - 1)} \left[\frac{Y\delta^{\delta}(1 - \delta)^{(1 - \delta)}}{M_a^{\delta} P_b^{(1 - \delta)}} \right]^{(\eta - 1)} \\
&\quad \times \left[\frac{(2\eta - 1)(\eta - 1)^{(\eta - 1)}}{\eta^{\eta}} - \frac{(2E - 1)(E - 1)^{(E - 1)}}{E^{E}} \right] \\
&= Q[f(\eta) - f(E)]
\end{aligned}$$

By using the series expansion of $\ln(1 - 1/\eta)$, it can be shown that $\partial f(\eta)/\delta\eta < 0$ Thus, since

$$Q > 0 \text{ and } E_{\sigma = 1} < \eta, \ \Delta W_{\substack{\sigma = 1 \\ e = g = \infty}} < 0$$

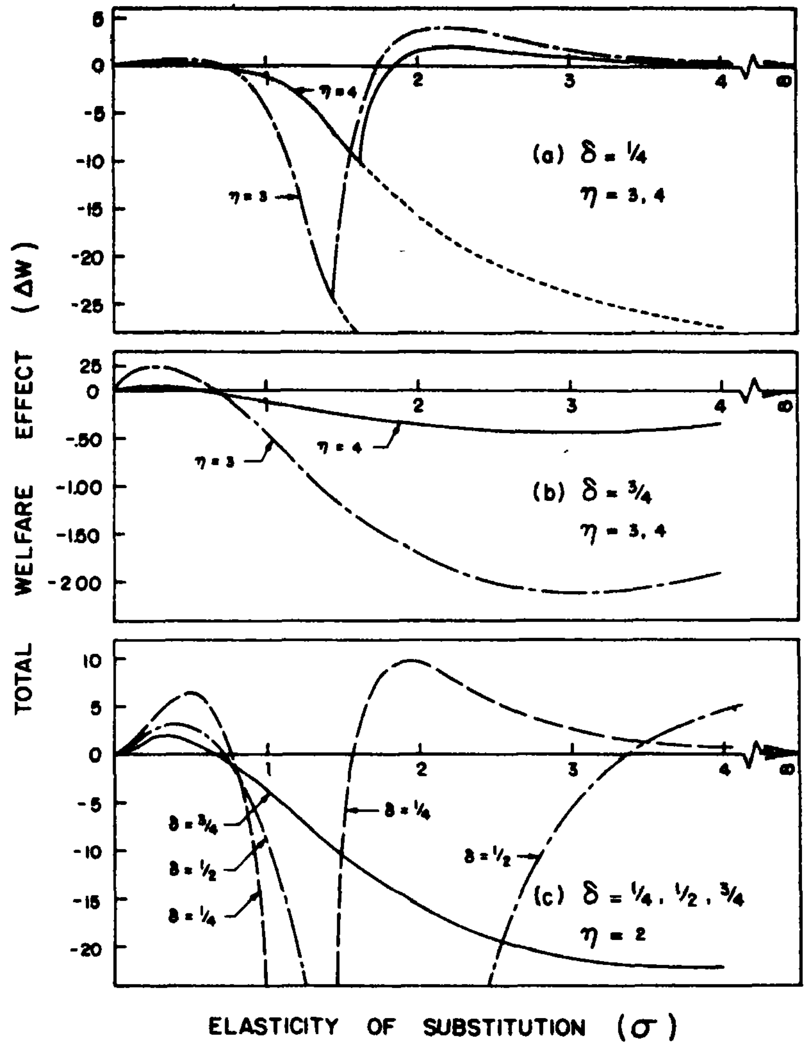


Fig. 4.—Total welfare effect of vertical control

set any desired price for the final product after vertical control, then the net welfare effect could be assumed negative at least for $\sigma \geq 1$ If, however, the input monopolist would be constrained from raising the price of the final product after vertical control to the full monopoly price, then the net welfare effect could not be assumed negative unless the elasticity of substitution is close to unity If σ is significantly less than unity and if the values for δ and η are relatively low, then the welfare effect might be presumed positive.

The result that the net welfare effect can be positive or negative is to be expected, since the welfare effect of an acquisition of a "second monopoly" is a priori indeterminate, given the presence of an existing monopoly distortion, and vertical control by an input monopolist can be viewed as the achievement of a "constrained second monopoly" over the other input industries Assuming constant costs in the input industries, a joint monopoly over all input industries is equivalent to a monopoly over the final product If the $\hat{P}_x$ constraint is not effective, then the same level of monopoly profits and the same price for the final product also result from vertical control by a single input monopolist, either in the form of vertical integration or by the imposition of a tying arrangement (the requirement that purchasers of A also purchase their requirements of B from the A monopolist) If the $\hat{P}_x$ constraint is effective, however, profits and final-product prices will be lower and total welfare higher under vertical control than under a joint monopoly of all inputs Under vertical control, the A monopolist sets the input prices P_a^* and P_b^* (either directly in a tying arrangement or implicitly through internal pricing under vertical integration) such that

$$\frac{P_a^*}{\bar{M}_a} = \frac{P_b^*}{\bar{P}_b} = \frac{\hat{P}_x}{\bar{M}_x} \leq \frac{\hat{P}_x}{\bar{M}_x} \qquad (12)$$

This "second monopoly" over the other input industries can be illustrated graphically if $\sigma = \eta$, since in this case the cross elasticity of demand between the inputs is zero and the input derived-demand curves do not shift in response to changes in the prices of other inputs If $\sigma = \eta$ (see fig 5), the effect of vertical control is simply to increase the price of the cooperant input while leaving the price of the monopolized input unchanged In figure 5*b*, $\Delta CS_x = -abcd$, $\Delta\pi = aecd$, and $\Delta W = -ebc$ If $\sigma > \eta$, then $E > \eta$ and $P_a^* > P_a$ Since $P_b^* > P_b$, the total welfare effect must therefore be negative if $\sigma \geq \eta$ and the $\hat{P}_x$ constraint is not effective If $\sigma < \eta$, $P_a^* < P_a$, and ΔW can be positive or negative Since $\eta > 1$, however, this conclusion is weaker than the simulation result that, if the $\hat{P}_x$ constraint is not effective, $\Delta W < 0$ if $\sigma \geq 1$

II. Vertical Control with Nonconstant Input Costs

If we relax the assumption that all input costs are constant, a fall in the price of the final product after vertical control becomes possible In addition, the effect of vertical control on producers' surplus in the competitive input industry must be considered

The effect on R of a rising or falling marginal-cost schedule for A can be determined from the effect of vertical control on the quantity of A used For example, if the quantity of A used increases after vertical control ($\Delta A > 0$), a rising M_a schedule will reduce ΔA and hence

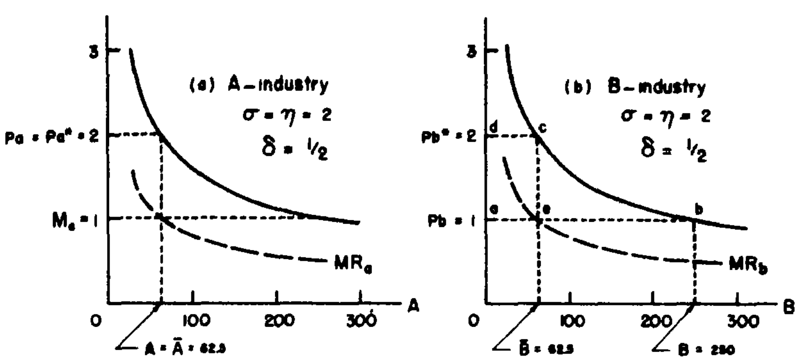


FIG 5 —Effect of vertical control on input industries for $\sigma = \eta$

increase R However, while the magnitude of ΔA will be affected by the form of the M_a schedule, the sign of ΔA will be the same whether M_a is rising, falling, or constant With constant M_a, the simulation results show that if the $\hat{P}_x$ constraint is not effective, the quantity of A used always falls after vertical control if $\sigma > \eta$, remains constant if $\sigma = \eta$, and increases if $\sigma < \eta$ If the $\hat{P}_x$ constraint is effective, however, the quantity of A used may increase even if $\sigma > \eta$ Thus if $\sigma < \eta$, then $\Delta A > 0$, and rising marginal costs in the monopolized input industry will increase R, while falling M_a will reduce R

Since the quantity of B used always falls after vertical control, the only way the price of the final product can decline after vertical control is if the quantity of A used increases sufficiently to offset the reduction in B The most favorable conditions for a large increase in A would be with $\sigma < \eta$[13] and a rapidly declining M_a If the elasticity of the marginal-cost schedule for A, g, is less than zero and small in absolute value, then $R < 1$ becomes possible For example, in the $\sigma = 1$ case, with constant costs in the B industry only ($e = \infty$, $g \neq \infty$), R reduces to

$$\underset{\substack{e=\infty \\ \sigma=1}}{R} = \left[\frac{(1 - 1/E)^{\delta g}}{(1 - 1/\eta)^{(g+1-\delta)}}\right]^{1/(g+E)}, \tag{13}$$

where

$$\underset{e=\infty}{E} = 1 + \delta(\eta - 1)$$

As expected, since $\sigma < \eta$,

$$\frac{\partial R}{\partial g} = \frac{\delta R}{(g + E)^2}\,[E \ln (1 - 1/E) - \eta \ln (1 - 1/\eta)] < 0$$

$$(\text{since } \sigma < \eta \rightarrow E < \eta)$$

---

[13] If the $\hat{P}_x$ constraint is effective, R must be greater than unity Thus, any increase in A if $\sigma \geq \eta$ cannot result in a fall in the price of the final product after vertical control

While the second-order conditions (if $g < 0$, then $g < -\eta$) restrict the range over which equation (13) can be less than unity, they do not exclude the possibility entirely The simulation results show a rather narrow range for g where the price of the final product will fall after vertical control. Thus if $\sigma < \eta$, a declining marginal-cost schedule clearly improves the welfare effect of vertical control (since the monopoly profits from vertical control are also increased) and may even result in a situation where vertical control improves the welfare of both consumers and profit receivers

A finite elasticity of supply of the competitive input ($\epsilon \neq \infty$) introduces a monopsony element In the absence of vertical control, the elasticity of supply of B enters indirectly through the elasticity of derived demand for A (the lower is ϵ, the lower is E in eq [4]) Vertical control, however, in addition to establishing the A monopolist as a "second monopolist" in the sale of B, now also enables the A monopolist to act directly as a monopsonist toward the B industry It is thus easy to understand why other input producers would object strenuously to vertical control by a firm with market power over an input with which their product can be used in variable proportions

From the viewpoint of consumers of the final product, however, a rising supply curve for B improves the effects of vertical control Since the quantity of B always falls after vertical control, a rising supply curve for B acts to reduce the fall in B after vertical control and hence reduce R If the supply of B is sufficiently inelastic, and if $\sigma < \eta$ so that $\Delta A > 0$, the price of the final product can fall after vertical control even if there are constant costs in the A industry For example, if $\sigma = 1$ and $\epsilon = 0$,[14]

$$\underset{\substack{\epsilon=0\\ \sigma=1}}{R} = \delta^{\delta g/\{\delta g+\eta[1+g(1-\delta)]\}} < 1 \tag{14}$$

Any such gain to consumers of the final product must, however, be balanced against the loss of producers' surplus in the B industry The total welfare effect of vertical control on consumers, monopoly profit receivers, and competitive input producers can, of course, be positive or negative depending on the particular parameter values For any given quantity of the final product, however, nonconstant costs in the B

---

[14]
$$\underset{\substack{\epsilon=0\\ \sigma=1}}{R} > 1$$
would require that $\delta g/\{\delta g + \eta[1 + g(1 - \delta)]\} < 0$, and thus $0 > g > -\eta/[\delta + \eta(1 - \delta)]$ or $0 > g > -E$, since

$$\underset{\substack{\epsilon=0\\ \sigma=1}}{E} = \frac{\eta}{\delta + \eta(1 - \delta)},$$

luded by the second-order conditions.

industry clearly reduce the potential efficiency gain from vertical control Before vertical control, the ratio of A to B is lower than the efficient ratio, but after vertical control the input ratio is higher than the efficient ratio if $\epsilon \neq \infty$:

$$\frac{\bar{A}}{\bar{B}} = \left[\frac{\bar{P}_b(1 + 1/\epsilon)}{(1 - \delta)\bar{M}_a}\right]^{\sigma} > \left[\frac{\delta \bar{P}_b}{(1 - \delta)\bar{M}_a}\right]^{\sigma}$$

III. Summary and Conclusions

This paper has examined the welfare effects of vertical control by a monopolist of an input which can be used in variable proportions with other inputs By specifying the production function (CES) and the demand function (constant elasticity) for the final product, the following conclusions can be drawn.

1 If the marginal costs of producing all inputs are constant, vertical control will always result in an increase in the price of the final product. If vertical control enables an unconstrained monopoly over the final product, the percentage increase in the price of the final product will be greater, the less elastic the demand for the final product, η, and the less "important" the monopolized input (the lower is the distribution parameter, δ, in the production function for the final product) However, if the final-product monopoly achieved through vertical control is effectively constrained by potential competition from firms which can produce the final product without using the monopolized input, the price increase will be less, the less elastic the final demand and the less important the monopolized input

2 If input costs are not constant, a fall in the price of the final product after vertical control is possible if (*a*) the elasticity of substitution between the inputs is less than the elasticity of demand for the final product, *and* either (*b*) the marginal cost schedule for the monopolized input is downward sloping, or (*c*) the supply curve of the competitive input is upward sloping

3 The combined welfare effect of vertical control on monopoly profit receivers, consumers of the final product, and producers of the competitive input may be positive or negative, depending on the particular parameter values This result is to be expected, since the welfare effect of a "second" monopoly (in this case, a monopoly over the other input industries) is a priori indeterminate, given the presence of an existing monopoly distortion With constant input costs, the combined welfare effect is always positive over a range for σ immediately above zero but becomes negative at some $\sigma < 1$ If the final-product monopoly achieved by vertical control is effectively constrained, the combined welfare effect becomes positive again at some $\sigma > 1$. In general, the

less elastic the demand for the final product and the less important the monopolized input, the greater the magnitude of the welfare effect and the wider the range for σ over which the combined welfare effect is positive

Further extension of these results would involve relaxing the assumptions made with respect to parameter values and the institutional context The critical parameter assumptions have been that σ and η are constant over the relevant range for vertical control The existence of a close substitute for A which could be used by independent X producers if P_a were increased slightly after vertical control would reduce P_x and hence R Similarly, a demand curve for the final product which is linear or even concave to the origin would reduce P_x and hence R Since the efficiency gain would still be possible, an increasing η or σ over the relevant range should improve the welfare effects of vertical control

The welfare effect would also clearly be improved if the A monopolist were unable to raise P_a significantly after vertical control This inability to raise the price of A facing independent X producers could be due to regulation, threat of antitrust proceedings, or potential entry at the A level In addition, if markets for A other than the X industry are present, an increase in P_a in these markets might be unavoidable if P_a is to be increased for independent X producers. If raising P_a in these other markets would reduce monopoly profits, the A monopolist may be unwilling to raise P_a to independent X firms after vertical control.

If the A monopolist is a labor union, rather than a firm, a change in the objective function is necessary Although a full analysis of the union case cannot be presented here, vertical control by unions, in a number of different forms,[15] permits the union to choose from a set of combinations of higher earnings per worker and higher employment levels The effect of vertical control by a union will then depend on the union's particular objective function For example, if the goal of the union is to maximize earnings per worker, subject to maintaining the existing level of employment, the effect of vertical control by the union will be to reduce the quantity of capital used in the industry while leaving the labor input constant. The result will be an increase in both earnings per worker and the price of the final product.

---

[15] First, the union could simply purchase the firms involved Second, the union could try to fix the labor/output ratio through work rules, featherbedding, or bargaining over both wages and the level of employment Third, the union could institute some form of a tying arrangement, increasing the marginal cost of capital relative to the marginal cost of labor through a profit-sharing agreement. (Unless the debt/equity ratio were fixed, however, effective vertical control would require that the profit-sharing agreement be based on net revenue before interest payments In addition, if inputs other than capital could be used in variable proportions with labor, a tying arrangement would have to cover these inputs as well) Fourth, the union could take its "monopoly profits" through a royalty on the final product, thus preventing any fall in the labor/output ratio resulting from unionization The best example of such a royalty arrangement is the United Mine Workers of America Welfare and Retirement Fund, which is financed by a 40-cent royalty on each ton of coal produced in unionized mines These royalty payments constitute a considerable share of total labor costs, reducing employment by significantly less than would a comparable financing system based on the level of employment

The sensitivity of the welfare effects of vertical control to the particular parameter values implies that accurate estimation of these parameters is necessary. In particular, however, empirical work on the elasticity of substitution has been concentrated on estimating σ between capital and labor in a value-added production function. These estimates fall mainly in the $0.5 < \sigma < 1.5$ range (see Nerlove 1967), where significant effects from vertical control on final product prices and monopoly profits could be expected (see figs. 1, 3). Estimates of the elasticity of substitution between intermediate inputs and value added are generally low.[16] Thus particular attention might be given to cases of vertical control by labor unions or capital good producers. An additional consideration is that the original basis of market power for unions and patented capital-good producers may be legally unassailable, while policies can still be formulated toward extensions of that market power through vertical control.

One implication of these estimates for σ is that, for practical purposes, the most relevant results in this paper are those for relatively low elasticities of substitution. The elasticity of substitution between one particular monopolized capital good or one category of labor and all other capital and labor inputs, however, can be expected to be higher than between capital and labor aggregates. Similarly, even if the elasticity of substitution between intermediate inputs and value added is negligible, substitution possibilities between different intermediate inputs may be significant. Thus we cannot rule out the intermediate-inputs case, nor can we ignore the results for high levels of σ.

References

Arrow, K., Chenery, H. B., Minhas, B., and Solow, R. S. "Capital-Labor Substitution and Economic Efficiency." *Rev. Econ. and Statis.* 43 (August 1961): 225–50.

Balassa, B., Guisinger, S. E., and Schydlowsky, D. M. "The Effective Rates of Protection and the Question of Labor Protection in the United States: A Comment." *J.P.E.* 78 (September/October 1970): 1150–62.

Burstein, M. L. "A Theory of Full-Line Forcing." *Northwestern Univ. Law Rev.* 55 (February 1960): 62–95.

Hicks, J. R. *The Theory of Wages.* New York: St. Martin's, 1963.

McKenzie, L. "Ideal Output and the Interdependence of Firms." *Econ. J.* 61 (December 1951): 785–803.

---

[16] See Theil and Tilanus (1964, p. 268) and Balassa, Guisinger, and Schydlowsky (1970, pp. 1157–61).

Mishan, E. J. "What Is Producer's Surplus?" *A.E.R.* 8 (December 1968): 1269–82.

Nerlove, M. "Recent Empirical Studies of the CES and Related Production Functions." In *The Theory and Empirical Analysis of Production*, edited by Murray Brown. NBER Studies in Income and Wealth, vol. 31. New York: Columbia Univ. Press, 1967.

Rader, T *Theory of Microeconomics* New York· Academic Press, 1972.

Schmalensee, R. "A Note on the Theory of Vertical Integration." *J.P.E* 81 (March/April 1973) 442–49.

Theil, H., and Tilanus, C B "The Demand for Production Factors and the Price Sensitivity of Input-Output Predictions" *Internat Econ Rev.* 5 (September 1964) 258–72

Vernon, J M., and Graham, D. A. "Profitability of Monopolization by Vertical Integration" *J.P E.* 79 (July/August 1971). 924–25

Williamson, O W "The Vertical Integration of Production Market Failure Considerations." *A E.R* 61 (May 1971) 112–23

Communications

The Risk Element in Occupational and Educational Choices: Comment

John C. Hause
University of Minnesota

In an interesting recent contribution, Weiss (1972) concludes that "moderate" degrees of risk aversion substantially reduce the average rates of return from educational investment. This conclusion depends critically on the maximand for the consumer assumed by Weiss, which is anti-Fisherian with a vengeance. The purpose of this note is to indicate more explicitly the role played by the utility functional specified by Weiss in leading to his conclusion, and to suggest an alternative procedure that may be preferred by economists addicted to traditional capital theory.

The two main results of the analysis are (*a*) that the original Weiss formulation is theoretically inappropriate for dealing with the problem at hand and (*b*) that the empirical result obtained by Weiss that moderate risk aversion greatly reduces the schooling returns to Ph.D. recipients is unjustified. That conclusion is primarily due to the foreseen absence of earnings while attending school, not to risk. The empirical discussion is also flawed by ignoring how the covariance structure of earnings over time must be taken into account in the empirical estimation of risk.

Weiss (1972, p. 1208) assumes that consumers attempt to maximize expected discounted utility

$$\text{Max } E\left\{\int_{t_0}^{T} e^{-\rho t}u[y(t)]\,dt\right\}, \tag{1}$$

Financial support for this note was provided by U.S. Office of Education grant OEG-0-72-1569. It has been modified at several points in response to correspondence with Yoram Weiss. Comments by Uri Ben-Zion were also useful. The author alone is responsible for reprehensibilities.

where current utility u is a function of current income $y(t)$, and has the form

$$u(y) = \frac{y^{1-a}}{1-a}, \tag{2}$$

where a is a constant.[1] Fisher's (1930) first principle of capital theory stresses the independence of consumption and production decisions under certainty (with a perfect capital market) and assumes utility is developed only from consumption. If one accepts this view, choosing equation (1) as the maximand can be justified only if intertemporal shifts in consumption through saving and borrowing are impossible. For a given discount rate, ρ, and for risk aversion ($0 < a < 1$), equation (1) implies that from the family of all income streams with the same present value, a known constant-level income stream is preferred, and is superior to receiving (with certainty) all income in the first or the last period of a person's working life, with zero incomes for other periods. This outcome follows from the assumed diminishing marginal utility of income at each point (or period) of time, not from risk, since the preceding comparison is between alternative income profiles received with certainty.[2] The postponement of income for 5 years past the B.A. degree by Ph.D. candidates assumed by Weiss, in conjunction with this feature, is the main element in leading to the reduced average rates of return he computes. Suppose that we compare two constant-level income streams, received with certainty, one extending from the B.A. to infinity, the second extending from the Ph.D. to infinity (with zero income for the 5 years between B.A. and Ph.D.). With a discount rate of $\rho = .10$, the Ph.D. income stream must have a present value between 1.13 and 7.39 times as large as the B.A. stream, as a varies between .2 and .8, to equalize utilities. This result has nothing to do with risk; it follows from zero earnings while acquiring the Ph.D. and the assumed intertemporal substitution implied in equation (1). Thus, it is not surprising that the coefficient of variation of income as a function of age (which presumably characterizes the actual risk[3]) is of trivial importance in reducing the average rates of return relative to the postponement of income for moderate a, a conclusion pointed out at the end of Weiss's study[4]

---

[1] Weiss acknowledges that the assumption of utility depending on current income is "somewhat dubious and is adopted only for simplicity" in n. 5 of his paper. However, his empirical conclusions depend critically upon this assumption. He suggests that one might view $u(y)$ as an indirect utility function, with a constant interest rate. This interpretation gets at the problem only for the trivial case where income profiles differ only by a scale parameter.

[2] Diminishing marginal utility of income was introduced in the expected utility hypothesis of von Neumann-Morgenstern to characterize risk aversion. It is technically misleading to conclude that the assumption of risk aversion necessarily implies anything about intertemporal preferences under certainty. This unsatisfactory formulation is obvious if we consider an individual with "risk preference" (by considering $-1 < a < 0$ in eq. [2]). In this case, the maximand of eq. (1) implies that of all certain income profiles with the same present value, the least desirable is a constant income over time, and the most desirable streams are those concentrated at one point (or period) of time.

For those accepting Fisherian capital theory, the natural maximand for getting at risk instead of income variability would be the expected utility of the discounted income stream[5] Of course, the probability distribution of discounted lifetime income cannot be computed from the cross-sectional data of scientists' incomes reported by Weiss, since it is impossible to derive the covariances of annual incomes over the life cycle

---

[3] Even the coefficient of variation of income as a function of age may be a misleading measure of uncertainty To the extent that there are substantial differences in postschool investment in on-the-job training anticipated by individuals, low initial incomes by those investing heavily in this way are offset by higher earnings later on

[4] Table 1 has kindly been made available by Weiss, and may be compared with table 2 in Weiss's original article to show that the strong reduction in average rates of return originally obtained there are primarily due to the zero incomes imputed to Ph D students for 5 years In this new table, Weiss used the reported incomes of scientists in educational institutions with B A and M A degrees as an estimate of student income (and presumably consumption) Thus, mean earnings in the new calculation are assumed to be $3,640, $4,467, $4,754, $5,041, and $5,328 per year instead of 5 years of zero earnings The coefficients of variation are 77 at the B A level and 67 at the M A level With a risk-aversion parameter of zero, this modification almost doubles the private industry average rate of return and more than triples the all-employers average rate of return, compared with table 2 in Weiss's article More significant, the average rates of return in this new table drop off at a relatively much slower rate as the risk-aversion parameter increases The range of the risk-aversion parameter in table 1 lies between zero and two, instead of the smaller range of zero to one in table 2 of Weiss's article The utility function used by Weiss decreases in value to negative infinity as income decreases to zero when the risk parameter exceeds one, and hence is not relevant for calculations assuming that Ph D students have zero earnings while they attend school The positive-earnings profile for Ph D students assumed by Weiss in calculating table 1 allows one to consider more extreme degrees of risk aversion than were considered in his original article For very high values of risk aversion (1 8 and 2), the computed rates of return are very small In both the original article and the new table, low rates of return are primarily a consequence of differences in level of the assumed mean-earnings profiles for those with B A and Ph D degrees These low returns thus depend on the intertemporal properties of the assumed utility functions, and not on anything related to risk

[5] The following discussion waives the problem of imperfect capital markets, which severely limit the possibility of people's obtaining consumption loans, to be paid off by subsequent earnings The original formulation by Weiss does not get at this complication either However, Weiss has correctly pointed out in subsequent correspondence that the sequential nature of consumption decisions makes the distribution of present values of income inadequate for characterizing the expected value of optimal consumption strategies I suspect that for empirical studies it may still turn out that the distribution of discounted lifetime earnings is an adequate entity for earnings-stream optimization It is clear that further theoretical work is needed to reveal under what circumstances and in what sense "adequacy" is attained In particular, there is currently little evidence on which specific utility functions of intertemporal consumption under uncertainty are most satisfactory for discussing these problems Furthermore, little seems to be known about how robust conclusions are which are based on calculations with an arbitrary utility function There may be a useful way of characterizing classes of utility functions and classes of earnings functions with random components such that the distribution of discounted earnings is suitable for analyzing schooling and occupational choice But this topic appears to be difficult

TABLE 1

AVERAGE RATES OF RETURN FOR VARIOUS DEGREES OF RISK AVERSION UNDER MODIFIED EARNINGS STREAM WHILE ATTENDING SCHOOL

	DEGREE OF RISK AVERSION										
TYPE OF EMPLOYMENT	0 0	0.2	0 4	0 6	0.8	1 0	1.2	1.4	1.6	1 8	2 0
Private industry	143	128	113	099	084	.069	055	040	028	.015	004
Government	180	160	140	121	103	086	069	.053	038	.024	011
All employers	177	162	146	131	.115	100	084	069	054	041	030

from a single cross section However, the upper bound of the coefficient of variation of discounted lifetime income (consistent with the observed standard deviation of income as a function of age) can be obtained, since the mean discounted value of lifetime income is the discounted value of the mean earnings profile, and the supremum of the standard deviation of discounted lifetime income is the discounted value of the standard deviation of income, as a function of age [6]

This calculation provides an upper bound to the coefficient of variation of a discounted income profile, consistent with the variances at each point along the profile The lower bound of the coefficient of variation is zero, which is approached if the income profile is determined by uncorrelated increments of income over short intervals of time These bounds are unpleasantly wide for estimating the true variability of discounted lifetime incomes, but greater precision requires information of income covariances along the income profile If data are limited to a single cross section, the estimating procedure developed by Fase (1971), based on a random-walk model, is probably superior to either of the bounding estimates

---

[6] I am indebted to Christopher Sims for pointing out an earlier error and sketching the correct derivation. The proof is shown below The corresponding result for variance is derived, and square roots may be taken of the first and last steps

$$\left\{\int e^{-\rho t}[E(y_t - \mu_t)^2]^{1/2}\,dt\right\}^2$$

$$= \iint \{E[e^{-\rho t}(y_t - \mu_t)]^2\}^{1/2}\{E[e^{-\rho t'}(y_t - \mu_t)]^2\}^{1/2}\,dt\,dt',$$

$$\geq \iint E[e^{-\rho t}(y_t - \mu_t)e^{-\rho t'}(y_{t'} - \mu_t)]\,dt\,dt',$$

by Schwarz's inequality, or equivalently, by the fact that the covariance of a pair of random variables is never greater than the product of the standard deviations of the variables

$$= E\iint e^{-\rho t}(y_t - \mu_t)e^{-\rho t'}(y_t - \mu_{t'})\,dt\,dt',$$

$$= E\left[\int e^{-\rho t}(y_t - \mu_t)dt\right]^2$$

References

Fase, M. M. G. "Estimation of Lifetime Income" *J American Statis. Assoc* 66, no. 336 (December 1971). 686–92.

Fisher, I. *The Theory of Interest* New York: Macmillan, 1930.

Weiss, Y. "The Risk Element in Occupational and Educational Choices." *J.P E* 80 (November/December 1972): 1203–13

Purchasing-Power Parity Theory and Economic Reform in Yugoslavia

Ryan C Amacher and John S Hodgson

University of Oklahoma

This study tests the applicability of purchasing-power parity theory to movements in the exchange rates between the Yugoslav dinar and two West European currencies during the 1950s and 1960s The purpose is to determine whether these exchange rates were administered without regard to underlying economic conditions, as is commonly thought, or whether they actually reflected changes in economic fundamentals within the selected countries

Although an absence of accessible, appropriate data precludes an examination of currency exchange rates and purchasing powers within Eastern Bloc countries, one of these Socialist countries, Yugoslavia, has been introducing market-oriented reforms more rapidly than other countries in Eastern Europe and has produced a set of economic statistics more readily comparable with those of Western countries These developments provide an opportunity to examine the exchange rates between this one East European currency and Western currencies [1] Accordingly, this study examines the exchange rates between the currency of Yugoslavia and those of her major Western trading partners, Italy and West Germany

The framework for analysis is the purchasing-power parity theory, which holds that the equilibrium exchange rate between two currencies is determined primarily by their relative purchasing powers [2] The theory became widely known following World War I, when it was formulated and expounded by Gustav Cassel Cassel originally contended that

---

We are grateful to Robert D Tollison and a referee for helpful comments

[1] Western observers have argued that exchange rates in Eastern Europe do not reflect economic forces and as such are irrational, irrelevant, artificial, unrealistic, and as characterized by one Eastern observer, basically wrong All quoted in Boltho (1971, p 64), "irrational," Brown (1962), "irrelevant," Pryor (1963), "artificial," Nove and Donnelly (1961), "unrealistic," United Nations For a more complete discussion of Western views on East European rates of exchange, see Boltho (1971) and Pryor (1963)

[2] For a discussion of purchasing-power parity theory, see Balassa (1964, pp 584–96), Yeager (1966, pp 170–88), and Holmes (1967, pp 686–95) For some empirical investigations relating to Western countries, see Graham (1930), Stolper (1948), Yeager (1958), Tsiang (1959), Gaillot (1970), and Hodgson (1972)

currency purchasing powers are the dominant long-run determinants of exchange rates Capital flows, transportation costs, and other trade impediments are of secondary importance. Subsequent writers have narrowed the theory to assert that relative currency purchasing powers, measured by relative price indexes, are the only important determinants of exchange rates, other factors being of negligible importance over the long run.[3] In its presently accepted form purchasing-power parity theory asserts that, if equilibrium is to be maintained, exchange rates must be kept in line with broad relative movements in international price levels, but that changes in such factors as tastes, resources, and technology will cause changes in exchange rates even at constant international prices.[4]

We find that the Yugoslav exchange rates are not independent of their purchasing powers, as represented by corresponding price levels within each of the countries This suggests that these exchange rates, widely thought to be arbitrary and unrelated to fundamentals, actually reflect underlying economic conditions within the three countries While the findings do not necessarily demonstrate that the Yugoslav authorities are wholly rational in their exchange rate policy, or that they are becoming more rational over time, they do suggest that the authorities are not out of touch with monetary fundamentals.

I. Purchasing-Power Parity Theory

The comparative version of purchasing-power parity theory states that the equilibrium exchange value of a currency in any period is equal to the value during a selected base period (in which the exchange rate is assumed to be in equilibrium) multiplied by the relative change in price levels in the two countries since the base year Thus

$$R_t = R_0 \; \frac{P^{*d}_t / P^{*d}_0}{P^{*f}_t / P^{*f}_0}, \qquad (1)$$

where R_t = the calculated purchasing-power parity exchange rate (in domestic currency units per unit of foreign currency) during period t,

R_0 = the actual exchange rate during the base period (also in domestic currency units per unit of foreign currency),

P^* = price level, f superscript denotes the foreign country, d superscript denotes the domestic country, 0 subscript denotes the base period, t subscript denotes the subsequent period

---

[3] Holmes 1967, pp 686–87

[4] However, as Yeager points out, "The doctrine does describe price-induced adjustments in trade flows that will limit non-monetary exchange-rate movement" (1958, pp. 518–19)

For any selected base period, R_0, P^{*d}_0, and P^{*f}_0 are constant, and formula (1) becomes

$$R_t = K \frac{P^{*d}_t}{P^{*f}_t}, \tag{1'}$$

where K is a constant.

The question of interest is how closely the actual exchange rate is associated with the calculated purchasing-power parity rate, or equivalently, how closely the actual rate is associated with the ratio of the price levels in the two countries We thus hypothesize the following relationship which will be used below as an estimating equation

$$\ln R(t) = \alpha_0 + \alpha_1 \ln P(t) + U(t) \qquad (\alpha'_1 > 0), \tag{2}$$

where R is the actual exchange rate in units of domestic currency per unit of foreign currency, and P is the ratio of the domestic to the foreign price index

II. The Yugoslav Experience

In 1947 Yugoslavia was the first country to adopt the Stalinist economic model and in 1953 became the first country to abandon it Since abandoning the Stalinist model and adopting her "own road to Socialism," Yugoslavia has followed an erratic course toward a market-planned economy This experiment has proceeded through a number of economic reforms Among other objectives, these measures were aimed at reforming the foreign trade sector and obtaining the eventual convertibility of the dinar[5] In recent years there have been several devaluations, necessary steps toward convertibility and participation in organized foreign exchange markets Still, these goals remain in the distant future[6] Yugoslavia has, nevertheless, become significantly involved in trade with the Western (non-Comecon) world, making plausible an application of a purchasing-power parity test to this administered rate of exchange[7]

III. Empirical Results

Estimates were made using end-year exchange rates and price indexes obtained from IMF sources for the period 1952–71[8] Since data on direct exchange rates between the dinar and the lira and between the

[5] See, for example, Anakioski (1969) and Horvat (1971)

[6] For an account of these reforms, see Amacher (1972) Although Yugoslavia still imposes exchange controls under a system of fixed rates and exchange rationing, and the country has attempted reform of the international sector, Yugoslav authorities give strong support to achieving convertibility of the dinar

[7] For measures of Yugoslavia's activity on Western and Eastern markets, see Amacher (1972, pp. 19–50)

[8] Wholesale price indexes were used for Italy and West Germany, and "producer's prices, industrial goods" were used for Yugoslavia Each of these series was obtained from *International Financial Statistics*.

dinar and the deutsche mark were unavailable, it was necessary to derive these from U S dollar exchange rates with these three currencies The "actual" dinar-lira exchange rate, R_{yl}, was thus derived as

$$R_{yl} = \frac{\text{number of dinar per dollar}}{\text{number of lira per dollar}},$$

and the "actual" dinar–deutsche mark exchange rate, R_{yg}, was computed as

$$R_{yg} = \frac{\text{number of dinar per dollar}}{\text{number of deutsche marks per dollar}}$$

Two series were available for the dinar-dollar exchange rate, one reflecting the "official" exchange rate, the other reflecting the "actual settlement rate " Both series were used since it would be interesting to compare the results based on the official exchange rates with those based on rates more closely related to market forces

The following notation is adopted to distinguish between these various exchange rates

R^1_{yl} = the dinar-lira exchange rate based on the official dinar-dollar rate,
R^2_{yl} = the dinar-lira exchange rate based on the actual settlement dinar-dollar rate,
R^1_{yg} = the dinar–deutsche mark exchange rate based on the official dinar-dollar rate,
R^2_{yg} = the dinar-deutsche mark exchange rate based on the actual settlement dinar-dollar rate

These computations implicitly assume consistent cross-rates between the currencies, an assumption that would not appear to be too unrealistic when dealing with annual data

The price-level ratios were computed directly using annual wholesale price indexes Thus P_{yl} = price index in Yugoslavia/price index in Italy, and P_{yg} = price index in Yugoslavia/price index in Germany

Equation (2) was estimated using generalized least squares (GLS) and the official exchange rate, yielding the following results [9]
for Italy·

$$\begin{array}{lll} \ln R^1_{yl}(t) = & -0\ 671 & +\ 3\ 543 \ln P_{yl}(t) \\ \text{SD} & (0\ 178) & (0\ 709) \end{array} \qquad (3)$$

$$r^2 = 45 \qquad N = 19 \qquad \text{D-W} = 1\ 52$$

for Germany

$$\begin{array}{lll} \ln R^1_{yg}(t) = & 4\ 328 & +\ 3\ 476 \ln P_{yg}(t) \\ \text{SD} & (0\ 199) & (0\ 736) \end{array} \qquad (4$$

$$r^2 = 41 \qquad N = 19 \qquad \text{D-W} = 1\ 69$$

---

[9] See Appendix for OLS estimates

In both cases relative movements in price levels are highly significant at the 05 level and explain between 40 and 45 percent of the variation in the exchange rates In each case a 1 percent change in relative price levels is associated with a 3 5 percent change in the corresponding exchange rate [10]

The results are also significant when actual settlement exchange rates are used, GLS estimates were as follows [11]
for Italy

$$\begin{array}{ll} \ln R^2_{yi}(t) = -0.341 + 3\ 031 \ln P_{yi}(t) & \\ \quad \text{SD} \qquad (0\ 121) \quad (0\ 516) & (5) \\ r^2 = 42 \qquad N = 19 \qquad \text{D-W} = 2\ 16 & \end{array}$$

for Germany

$$\begin{array}{ll} \ln R^2_{yg}(t) = 4\ 676 + 2\ 846 \ln P_{yg}(t) & \\ \quad \text{SD} \qquad (0\ 135) \quad (0\ 507) & (6) \\ r^2 = 39 \qquad N = 19 \qquad \text{D-W} = 2\ 18 & \end{array}$$

Relative price-level movements are again highly significant at the 05 level and explain roughly 40 percent of the variation in the exchange rates In each case a 1 percent change in relative price levels is associated with roughly a 3 percent change in the exchange rates As would be expected, the price-level coefficients suggest that the actual settlement rate was adjusted slightly more in line with movements in price levels than was the more arbitrary official rate

Since structural changes are bound to have occurred within the three economies over the 20-year period under study, it would have been desirable to include time variables in the regressions to capture these effects Unfortunately the price-level ratios are highly correlated with time for each pair of countries, and the inclusion of time variables would introduce strong elements of multicollinearity [12]

In each case the observed relationships between exchange rates and price levels are weakened by the fact that the initial exchange rate is not likely to have been an equilibrium rate. Any movement of the actual exchange rate during subsequent periods, from an initial disequilibrium level toward an underlying equilibrium level, would appear as a movement unrelated to variations in price levels during the subsequent periods Since initial disequilibrium was probably present, the r^2s in these estimates are probably biased in a downward direction

---

[10] This might be explained by the fact that as Yugoslav authorities attempt to establish convertibility, a goal of the reform, the exchange rate is being adjusted not only to account for recent relative price changes, but also to adjust for past relative price changes that had not previously been incorporated in the administered rate of exchange

[11] See Appendix for OLS estimates

[12] Regressions including a time variable are nevertheless reported in the Appendix

IV. Conclusion

The results support the hypothesis that the devaluations in the dinar's rate of exchange during the 1950s and 1960s conform to patterns predicted by purchasing-power parity theory Despite all the problems of measurement,[13] it appears that Yugoslavia's prices and exchange rates have not been administered as independently of economic fundamentals as has been generally believed While these findings should not be loosely generalized to apply to intra-Bloc relationships, they do raise the suspicion that, in some of the other Socialist economies, exchange rate levels may not be totally "irrational, irrelevant, artificial, or unrealistic," but may, in fact, be responsive to basic economic fundamentals [14] More definite answers await more complete data and more comprehensive tests.

Appendix

I Ordinary least squares estimates

A Official exchange rate

For Italy

$$(3') \quad \ln R^{1}_{yi}(t) = -0.841 + 4\ 614 \ln P_{yi}(t)$$
$$\text{SD} \qquad (0.127) \quad (0\ 584)$$
$$r^2 = 78 \qquad N = 20 \qquad D\text{-}W = 0\ 57$$

For Germany

$$(4') \quad \ln R^{1}_{yg}(t) = 4\ 240 + 3\ 941 \ln P_{yg}(t)$$
$$\text{SD} \qquad (0\ 125) \quad (0\ 493)$$
$$r^2 = 78 \qquad N = 20 \qquad D\text{-}W = 0\ 57$$

B Actual settlement rate

For Italy

$$(5') \quad \ln R^{2}_{yi}(t) = -0\ 319 + 3\ 125 \ln P_{yi}(t)$$
$$\text{SD} \qquad (0\ 089) \quad (0\ 408)$$
$$r^2 = 77 \qquad N = 20 \qquad D\text{-}W = 1.01$$

For Germany

$$(6') \quad \ln R^{2}_{yg}(t) = 4\ 740 + 2\ 724 \ln P_{yg}(t)$$
$$\text{SD} \qquad (0\ 090) \quad (0\ 356)$$
$$r^2 = 76 \qquad N = 20 \qquad D\text{-}W = 0.95$$

---

[13] For a discussion of some of these problems, see Wiles (1969, pp 123–55)

[14] Gunter Kohlmey, the East German National Prize winning economist, argues th case both ways in a single volume. First, "Exchange rates are stable and are bein increasingly brought into alignment with the purchasing power of single currency units.' Later in the same volume he asserts that the exchange rates do not really correspond t the purchasing power of different currencies (see Pryor 1963, p. 238).

II. **Generalized least squares estimates using time variables (T)**

A. **Official exchange rate**

For Italy:

$$\begin{array}{llll} (3'')\ \ln R^1_{yi}(t) = & -4.162 + & 1.797 \ln P_{yi}(t) + & 0.062\,T \\ \quad\text{SD} & (1.493) & (1.075) & (0.027) \end{array}$$

$$r^2 = .70 \qquad N = 19 \qquad D\text{-}W = 1.73$$

For Germany.

$$\begin{array}{llll} (4'')\ \ln R^1_{yg}(t) = & -0.076 + & 1.234 \ln P_{yg}(t) + & 0.079\,T \\ \quad\text{SD} & (1.764) & (1.137) & (0.032) \end{array}$$

$$r^2 = .74 \qquad N = 19 \qquad D\text{-}W = 1.81$$

B. Actual settlement rate

For Italy.

$$\begin{array}{llll} (5'')\ \ln R^2_{yi}(t) = & -3.559 + & 0.849 \ln P_{yi}(t) + & 0.059\,T \\ \quad\text{SD} & (1.056) & (0.823) & (0.019) \end{array}$$

$$r^2 = .71 \qquad N = 19 \qquad D\text{-}W = 2.03$$

For Germany

$$\begin{array}{llll} (6'')\ \ln R^2_{yg}(t) = & 0.689 + & 0.420 \ln P_{yg}(t) + & 0.073\,T \\ \quad\text{SD} & (1.117) & (0.724) & (0.020) \end{array}$$

$$r^2 = .78 \qquad N = 19 \qquad D\text{-}W = 1.95$$

When time is included in the double-log specification, the t-scores for the coefficients of the relative price level variables drop below 2.0, while those for the time variable remain above 2.0. In nonlog form (not reported), the t-scores for the relative price level variables remain above 2.0 while those for the time variable fall. This illustrates problems caused by the presence of multicollinearity when the time variable is included.

References

Amacher, R. *Yugoslavia's Foreign Trade: A Study of State Trade Discrimination.* New York: Praeger, 1972.

Anakioski, D. "Foreign Trade in the Yugoslav Reform." *Yugoslav Survey* (1969).

Balassa, B. "The Purchasing-Power Parity Doctrine: A Reappraisal." *J.P.E.* (December 1964).

Boltho, A. *Foreign Trade Criteria in Socialist Economies.* Cambridge: Cambridge Univ. Press, 1971.

Brown, A. "Centrally Planned Foreign Trade and Economic Efficiency." *American Econ.* (November 1962).

Gailliot, H. "Purchasing-Power Parity as an Explanation of Long Term Changes in Exchange Rates." *J. Money, Credit, and Banking* (August 1970).

Graham, F. *Exchange, Prices, and Production in Hyper-Inflation: Germany 1920–1923.* Princeton, N.J.: Princeton Univ. Press, 1930.

Hodgson, J. "An Analysis of Floating Exchange Rates: The Dollar-Sterling Rate, 1919–1925." *Southern Econ. J.* (October 1972).

Holmes, J. "The Purchasing Power Parity Theory: In Defense of Gustav Cassel as a Modern Theorist." *J.P.E.* (October 1967).

Horvat, B "Yugoslav Economic Policy in the Post-War Period: Problems, Ideas, Institutional Developments." *American Econ Rev.* (suppl., June 1971).

International Monetary Fund *International Financial Statistics* Washington: IMF, selected years

Nove, A, and Donnelly, D *Trade with the Communist Countries* London Hutchinson, 1961

Pryor, F "Foreign Trade Theory in the Communist Bloc" *Soviet Studies* (July 1962)

——— *The Communist Foreign Trade System* Cambridge Cambridge Univ Press, 1963

Stolper, W F "Purchasing-Power Parity and the Pound Sterling from 1919-1925" *Kyklos*, fasc 3 (1948)

Tsiang, S "Fluctuating Exchange Rates in Countries with Relatively Stable Economies Some European Experiences after World War I" *IMF Staff Papers* (October 1959)

United Nations, Economic Commission for Europe *Econ Bull Europe* (November 1967)

Wiles, P *Communist International Economics* New York Praeger, 1969

Yeager, L "A Rehabilitation of Purchasing-Power Parity" *J P E* (December 1958)

——— *International Monetary Relations* New York Harper & Row, 1966

Authoritarian versus Nonauthoritarian Approaches to Economic Development

G William Dick

International Aero Service Corporation

A frequently stated and widely believed maxim of economic development holds that strong central planning under a communist or authoritarian regime is necessary for an underdeveloped country to achieve rapid economic growth Robert Heilbroner (Heilbroner 1963, pp 132, 135, and 138), for example, states that "only political leadership of the most forceful kind can carry the Great Ascent along " He goes on to say that "in most of the underveloped nations the choice for the command post of development is apt to lie between a military dictatorship and a left-wing civilian dictatorship the logic of events points to the formation of economic systems and political regimes which will seek to *impose* development on their peoples Communism may well be the quickest possible way out of underdevelopment, and the desperate fact is that in many areas of the world the present non-communist effort looks like the slowest possible way "

Jagdish Bhagwati (Bhagwati 1966, pp 203–4) trumpets the same refrain even more forcefully

> No policy of economic development can be carried out unless the government has the capacity to adhere to it Quite often, however, democratic governments lose equanimity and determination in the face of opposition This is the dilemma of most democratic governments It is here that socialist countries have an immense advantage their totalitarian structure shields the government from the rigorous and reactionary judgments of the electorate Another advantage of the socialist countries is their passionate conviction and dedication to the objective of economic growth—which contrasts visibly with the halting and hesitant beliefs and actions of most democracies The firm and purposive sense of direction is in pointed contrast to the extensive revisions and changes in policies and

Appreciation is expressed to Ernest A Duff, professor of politics, Randolph-Macon Woman's College, for his assistance in classifying the independent less developed countries of the world by form of government

> methods which are prompted by minor setbacks in most democratic governments and which produce a sense of drift and helplessness. The political economy of development poses, in this respect, a cruel choice between rapid (self-sustained) expansion and democratic processes

This theme is so pervasive in the literature that it is generally accepted as fact, though the assertion is rarely supported by definitive data of any kind. Recently a group of college students was asked, toward the end of a seminar course on economic and political development in which they had been exposed to an extensive range of literature in the field, whether they thought an authoritarian or a competitive form of government was more conducive to economic development The reply was virtually unanimous —authoritarian

Such unquestioning acceptance of an unsupported precept prompted an analysis of this question Is there any meaningful relationship between the form of government in a less developed country (LDC) and its rate of economic growth?

The independent LDCs of the world were classified as to form of government authoritarian, semicompetitive, or competitive An authoritarian regime is defined as a government where one political party maintains control through force and/or propaganda, or where there is no political party, with control maintained by one or more persons Elections, if they are held, are meaningless

A semicompetitive government means that there is either one major and several minor political parties, with the major party winning all elections, or one political party which conducts legitimizing elections from time to time Also, control over the population is not extensive

A competitive form of government is one where two or more political parties are contending for power, with free elections held at some regular interval

Classification reflects the general political state of a country during the 1960s, rather than the state of the country at any point in time This may result, for example, in a classification of a country as semicompetitive, though it may presently have a competitive political system

These countries were further subdivided by region of the world—Africa, America, Asia, and Europe. This classification scheme for the less developed countries studied is shown in Appendix A. All independent LDCs of the world were considered for analysis, and only India and those countries for which reliable economic data were not available were excluded. India was not included simply because its huge size would obscure the data on average rates of growth for the other LDCs of the world Mainland China would have been excluded for the same reason, but tha choice was not open because no data were available for mainland Chin

TABLE 1

LDCs RANKED BY REAL GDP PER CAPITA GROWTH RATES (1959–68)
BY FORM OF GOVERNMENT
(QUARTILE DISTRIBUTION IN %)

QUARTILE	FORM OF GOVERNMENT Authoritarian	 Semicompetitive	 Competitive
1	30	21	23
2	15	42	23
3	26	16	38
4	30	21	15

SOURCE —Derived from Appendix B

in any case Thus, 58 countries, plus the Francophone countries (consisting of 14 African nations) were included in this study. Data on the Francophone countries as a group were available, though data on the individual countries were not These countries, as a group, were classified as semicompetitive, by weighting the 14 individual forms of government by population

The period covered by this study is 1959–68, with at least 5 years of independence and reliable economic data during that period required for inclusion of a country Thus, a LDC which achieved independence in 1964 would be included, while one gaining independence in 1965 would not

Data on real gross domestic product (GDP) and population during the 1960s were collected and the average annual rate of growth of real GDP and real GDP per capita were calculated for each country. The growth rates calculated are trend rates, the trend being determined by fitting a semilogarithmic straight line to the data by the method of least squares The LDCs studied are ranked by real GDP per capita growth rates in Appendix B and by real GDP growth rates in Appendix C Table 1 summarizes Appendix B, showing the quartile distribution of each form of government when ranked according to real GDP per capita growth rates.

Thirty percent of the LDCs having authoritarian regimes placed in the top quartile, a larger percentage than that attained by either semicompetitive (21 percent) or competitive (23 percent) governments. But 30 percent of the authoritarian countries also fell in the bottom quartile, twice the proportion of competitive countries ranking this low, and 56 percent fell below the median rank. Only 37 percent of the semicompetitive countries placed below the median rank, though 53 percent of the competitive countries did so Forty-one percent of the authoritarian countries fell in the interquartile range, with 30 percent above and 30 percent below this range Sixty-one percent of the competitive governments were in the interquartile range, with 23 percent above and only 15 percent below this range. This suggests that LDCs with authoritarian forms of government perform either very well or very poorly Though, in general,

TABLE 2

LDCs Ranked by Real GDP Growth Rates (1959–68)
by Form of Government
(Quartile Distribution in %)

Quartile	Form of Government Authoritarian	Semicompetitive	Competitive
1	26	21	31
2	22	37	15
3	22	21	38
4	30	21	15

Source.—Derived from Appendix C.

competitive governments may not be able to achieve the high growth rates attained by authoritarian countries, neither are their chances of failure so high, and the overall performance of semicompetitive governments generally surpasses that of either of the other two forms of government.

Table 2 summarizes Appendix C, showing the quartile distribution of each form of government when ranked according to real GDP growth rates. Much the same conclusion would be drawn from this table, leaning even more in favor of competitive forms of government as opposed to authoritarian regimes with respect to real GDP growth rates. Competitive governments placed 31 percent in the top quartile, surpassing both authoritarian (26 percent) and semicompetitive (21 percent) countries, while only 15 percent fell in the bottom quartile.

Ranked either by real GDP per capita growth rates or by real GDP growth rates, the mean ranks of competitive, semicompetitive, and authoritarian countries were 28, 30, and 31, respectively. This also seems to favor a nonauthoritarian, rather than an authoritarian, approach to economic development.

In analyzing this question further, an average annual rate of growth, weighted by population, was calculated for each group of countries, according to the above classification scheme. These results are shown in table 3.

These figures reveal that the 27 authoritarian LDCs and the 13 competitive LDCs of the world achieved exactly the same rates of growth i real GDP per capita (1.4 percent per year) and real GDP (2.9 percent pe year) over the 1959–68 period. The 19 semicompetitive LDCs of th world (one of which is the group of 14 Francophone countries) achieve slightly higher growth rates of 1.6 percent and 3.5 percent per year in re GDP per capita and real GDP, respectively.

The LDCs with a competitive form of government attained higher rat of economic growth (real GDP per capita) than those with an authorita

TABLE 3

REAL GDP PER CAPITA GROWTH RATES (%/YR), BY FORM OF GOVERNMENT AND BY REGION, 1959–68*

REGION	FORM OF GOVERNMENT			
	Authoritarian	Semicompetitive	Competitive	All
Africa	1.7 (3.5)	1.5 (3.5)	2.0 (2.8)	1.6 (3.5)
America	0.7 (2.2)	1.4 (3.0)	1.8 (3.4)	1.1 (2.6)
Asia	1.2 (2.9)	2.4 (3.9)	0.9 (2.5)	1.2 (2.9)
Europe	3.4 (3.9)	3.2 (3.6)	1.5 (2.7)	2.8 (3.5)
All	1.4 (2.9)	1.6 (3.5)	1.4 (2.9)	1.5 (3.1)

*Average rates, weighted by population. Real GDP growth rates are shown in parentheses.

ian or semicompetitive form of government in Africa and America. The authoritarian governments in Europe showed the best performance, while in Asia they performed slightly better than the competitive governments but not nearly as well as the semicompetitive governments.

By region, the fastest economic growth was attained in Europe and the slowest in America and Asia, while the growth rates achieved in Africa fell in between, very close to the world average. These regional comparisons must be considered of limited significance in some cases, due to the numbers of countries involved. Only six of the countries studied are located in Europe, three authoritarian, two competitive, and one semicompetitive. Furthermore, there are only three semicompetitive and two competitive countries in Asia and only one competitive political system in Africa.

Though India, with a competitive form of government, was excluded from this study, a brief look at the Indian experience is warranted. From 1959 to 1968, real GDP per capita in India grew at an average annual rate of only 0.5 percent, considerably below the world average. The rate of growth in real GDP was 1.7 percent, again somewhat below the world average.

Likewise, a glance at four less developed communist countries in Europe (Albania, Bulgaria, Poland, and Rumania), which were excluded due to a lack of reliable economic data, is informative. Based on the limited data available, Rumania achieved the fastest rate of growth of these four countries, 8.2 percent per year in real net material product (NMP) per capita and 9.0 percent per year in real NMP, while Poland showed the slowest gains, 2.3 percent and 2.9 percent. The average annual rates of growth for these four countries taken together (weighted by population),

TABLE 4

REAL GDP PER CAPITA GROWTH RATES (%/YR), BY FORM OF GOVERNMENT AND BY REGION, 1959–68*

REGION	FORM OF GOVERNMENT			
	Authoritarian	Semicompetitive	Competitive	All
Africa	3.6 (5.6)	2.1 (4.2)	2.0 (2.8)	2.6 (4.6)
America	0.7 (1.8)	1.4 (3.1)	3.9 (5.2)	2.3 (3.7)
Asia	1.7 (3.4)	2.3 (3.9)	0.9 (2.5)	1.7 (3.3)
Europe	3.3 (3.8)	3.2 (3.6)	1.6 (2.8)	3.0 (3.6)
All . . .	2.0 (3.2)	1.9 (3.6)	3.3 (4.7)	2.3 (3.7)

*Average rates, weighted by GDP. Real GDP growth rates are shown in parentheses.

from 1960 to 1968, were 4.9 percent in real NMP per capita and 5.7 percent in real NMP. These rates are notably higher than the world average rates of growth for LDCs.

Determining the average rates of growth, weighted by GDP rather than by population, alters these conclusions slightly—and in favor of competitive forms of government. These results are shown in table 4.

The most notable change resulting from the use of GDP rather than population weights is the higher rates of growth in competitive countries relative to both authoritarian and semicompetitive countries. Weighted by population, competitive and authoritarian countries showed the same rates of growth, both slightly below the growth rates of semicompetitive countries. Now, weighted by GDP, countries with competitive systems of government exhibit growth rates of 3.3 percent per year in real GDP per capita and 4.7 percent per year in real GDP—both significantly higher than the rates attained by either authoritarian governments (2.0 percent and 3.2 percent) or semicompetitive governments (1.9 percent and 3.6 percent).

Proponents of an authoritarian approach to economic development generally hold that this is especially necessary during the earliest stages of development. Since this contention, again, is rarely supported by any data, the present analysis was extended to shed some light on its validity. The median GDP per capita, at the beginning of the time period studied, for the 59 LDCs analyzed was determined. Then the average rates of growth for countries above and below this median value, by form of government, were calculated. These results are shown in table 5.

When the individual rates are weighted by population in arriving at the average rates of growth, the data clearly do not support the argument tha

TABLE 5

REAL GDP PER CAPITA GROWTH RATES (%/YR), BY FORM OF GOVERNMENT, ABOVE AND BELOW INITIAL MEDIAN GDP PER CAPITA, 1959–68*

RELATION TO MEDIAN	FORM OF GOVERNMENT: Authoritarian	Semicompetitive	Competitive	All
	Averages (Weighted by Population)			
Above	2.6 (3.6)	1.9 (3.6)	1.7 (3.1)	2.1 (3.5)
Below	1.0 (2.7)	1.5 (3.4)	1.0 (2.5)	1.2 (2.9)
	Averages (Weighted by GDP)			
Above	2.4 (3.4)	2.0 (3.7)	3.6 (4.9)	2.7 (4.0)
Below	1.3 (2.9)	1.5 (3.3)	1.0 (2.6)	1.3 (3.0)

*Real GDP growth rates in percentage change per year are shown in parentheses.

authoritarian governments perform better than competitive governments do in the earliest stages of development. Below the median GDP per capita, the authoritarian countries exhibited the same rate of growth in real GDP per capita as the competitive countries did, and only a slightly higher rate in real GDP. The countries with semicompetitive forms of government experienced higher growth rates than either of the other two groups of countries. Above the median, authoritarian governments resulted in slightly faster rates of growth than either competitive or semicompetitive governments. This suggests that authoritarian governments do not perform better than nonauthoritarian governments in the earliest stages of economic development, though they may surpass these other forms of government in later stages of development.

When the individual rates are weighted by GDP, the data do tend to support this argument. The authoritarian countries perform better than the competitive countries below the median GDP per capita, whereas the reverse is true above the median. Yet it must be noted that, again, the LDCs below the median GDP per capita with semicompetitive forms of government outperformed both the authoritarian and competitive countries. Once more, this lends considerable doubt to the flat statement that the best approach to economic development is one with an authoritarian government.

In summary, these results certainly do not support, and tend to refute, the view that authoritarian countries are universally capable of achieving faster economic growth in the early stages of development than countries having competitive political systems, or that underdeveloped countries face "a cruel choice between rapid (self-sustained) expansion and democratic processes" (Bhagwati 1966, p 204). The weight of the evidence presented here suggests that the opposite presumption is more valid, requiring exponents of this view to support such a contention The evidence indicates, at the very least, that the choice is not at all a clear-cut one and depends on the circumstances prevailing in each country

Appendix A

LESS DEVELOPED COUNTRIES INCLUDED IN STUDY, BY FORM OF GOVERNMENT AND BY REGION

	FORM OF GOVERNMENT		
REGION	Authoritarian	Semicompetitive	Competitive
Africa	Egypt Ghana Libya Morocco Southern Rhodesia Uganda Zaire	Algeria Ethiopia Kenya Nigeria South Africa Sudan Tanzania Tunisia Francophone countries*	Sierra Leone
America	Argentina Bolivia Brazil Cuba Haiti Honduras Paraguay	Dominican Republic Ecuador Guatemala Mexico Nicaragua Peru	Chile Colombia Costa Rica El Salvador Jamaica Panama Trinidad and Tobago Uruguay
Asia	Burma China (Taiwan) Indonesia Iraq Khmer Republic Pakistan South Korea South Vietnam Syria Thailand	Iran Jordon Malaysia	Ceylon Philippines
Europe	Portugal Spain Yugoslavia	Greece	Cyprus Turkey

SOURCE —Based principally on *Political Handbook and Atlas of the World, 1970* (Stebbins and Amoia 1970)

*The Francophone countries include Cameroon, Central African Republic, Chad, Congo, Dahomey Gabon, Ivory Coast, Madagascar, Mali, Mauritania, Niger Senegal, Togo, and Upper Volta

Appendix B

LESS DEVELOPED COUNTRIES OF THE WORLD, RANKED ACCORDING TO REAL GDP PER CAPITA GROWTH RATES, 1959–68*

Rank	Country and Form of Government†		Rate (%/yr)
1	Libya, 1962–68‡	A	22.7
2	Jordan, 1959–67	S	6.6
3	Portugal, 1960–68	A	5.0
4	Cyprus, 1960–68	C	5.0
5	Chile	C	4.8
6	South Africa, 1960–68	S	3.9
7	China (Taiwan)	A	3.3
8	Greece	S	3.2
9	Spain	A	3.2
10	Yugoslavia	A	2.8
11	South Vietnam, 1960–66	A	2.5
12	Jamaica, 1962–68‡	C	2.5
13	South Korea	A	2.4
14	Iran	S	2.4
15	Uganda, 1962–68‡	A	2.4
16	Panama	C	2.3
17	Thailand	A	2.2
18	Nigeria, 1960–66‡	S	2.1
19	Trinidad and Tobago, 1962–66	C	2.0
20	Zaire, 1960–64	A	2.0
21	Siera Leone, 1963–68	C	2.0
22	Iraq, 1959–66	A	2.0
23	Nicaragua	S	1.9
24	Syria, 1959–67§	A	1.9
25	Tanzania, 1964–68	S	1.7
26	Mexico	S	1.6
27	Malaysia, 1963–68§	S	1.6
28	Kenya, 1963–68	S	1.6
29	Francophone, 1960–68	S	1.5
30	Peru	S	1.4
31	Turkey	C	1.4
32	Ethiopia	S	1.4
33	Egypt 1959–66‖	A	1.4
34	Pakistan‡	A	1.4
35	El Salvador	C	1.4
36	Bolivia	A	1.3
37	Tunisia	S	1.1
38	Costa Rica	C	1.0
39	Philippines	C	0.9
40	Guatemala	S	0.9
41	Khmer Republic, 1959–64	A	0.9
42	Ceylon‡	C	0.9
43	Honduras	A	0.8
44	Brazil	A	0.8
45	Argentina	A	0.8
46	Burma	A	0.8
47	Colombia	C	0.8
48	Ecuador	S	0.6
49	Morocco#	A	0.4
50	Paraguay	A	0.4
51	Southern Rhodesia, 1959–67	A	0.2
52	Ghana	A	0.1
53	Cuba, 1962–66**	A	0.0
54	Sudan, 1959–66	S	−0.2
55	Indonesia, 1959–67	A	−0.2
56	Haiti, 1959–64‡	A	−0.3
57	Dominican Republic	S	−0.3
58	Uruguay	C	−0.3
59	Algeria, 1959–66#	S	−2.6

SOURCE.—Derived principally from *National Accounts of Less Developed Countries, 1959–1968* (Organization For Economic Cooperation and Development 1970). Also, *Yearbook of National Accounts Statistics, 1969* (United Nations 1970) and *Demographic Yearbook, 1970* (United Nations 1971).

*Unless otherwise indicated, data are based on GDP at constant market prices, in national currency for the period 1959–68.

†A = authoritarian, S = semicompetitive, and C = competitive.

‡Based on GDP at constant factor costs.

§Based on GNP at constant market prices.

‖Based on net domestic product at constant factor costs.

#Based on gross domestic "production" at constant market prices. Contrary to GDP, it excludes by definition value added by civil servants, household servants, private nonprofit institutions, and financial intermediaries.

**Based on gross material product at constant market prices.

Appendix C

Less Developed Countries of the World, Ranked according to Real GDP Growth Rates, 1959–68*

Rank	Country and Form of Government†		Rate (%/yr)	Rank	Country and Form of Government†		Rate (%/yr)
1	Libya, 1962–68‡	A	27.2	31	Peru	S	3.0
2	Jordan, 1959–67	S	9.6	32	Sierra Leone, 1963–68	C	2.8
3	South Africa, 1960–68	S	6.4	33	Costa Rica	C	2.7
4	Chile	C	6.1	34	Pakistan‡	A	2.7
5	Cyprus, 1960–68	C	6.0	35	Egypt, 1959–66‖	A	2.7
6	Portugal, 1960–68	A	6.0	36	Turkey	C	2.7
7	South Vietnam, [illegible]0–66	A	5.3	37	Philippines	C	2.7
8	Jamaica, 1962–68‡	C	5.0	38	Bolivia	A	2.6
9	Uganda, 1962–68‡	A	4.9	39	Cuba, 1962–66**	A	2.5
10	China (Taiwan)	A	4.9	40	Honduras	A	2.5
11	Syria, 1959–67§	A	4.9	41	Guatemala	S	2.5
12	Trinidad and Tobago, 1962–66	C	4.6	42	Colombia	C	2.4
13	Nigeria, 1960–66‡	S	4.5	43	Ethiopia	S	2.3
14	Zaire, 1960–64	A	4.3	44	Brazil	A	2.3
15	Tanzania, 1964–68	S	4.2	45	Ecuador	S	2.3
16	Panama	C	3.9	46	Tunisia	S	2.2
17	Iran	S	3.8	47	Khmer Republic, 1959–64	A	2.2
18	South Korea	A	3.8	48	Indonesia, 1959–67‡	A	2.2
19	Thailand	A	3.7	49	Ceylon‡	C	2.1
20	Francophone, 1960–68	S	3.7	50	Paraguay	A	1.9
21	Spain	A	3.6	51	Burma	A	1.8
22	Nicaragua	S	3.6	52	Morocco#	A	1.8
23	Greece	S	3.6	53	Argentina	A	1.6
24	Iraq, 1959–66	A	3.6	54	Dominican Republic	S	1.5
25	Southern Rhodesia, 1959–67	A	3.4	55	Ghana	A	1.4
26	Yugoslavia	A	3.4	56	Sudan, 1959–66	S	1.3
27	Mexico	S	3.3	57	Haiti, 1959–64‡	A	0.7
28	El Salvador	C	3.2	58	Uruguay	C	0.3
29	Malaysia, 1963–68§	S	3.1	59	Algeria, 1959–66#	S	−1.5
30	Kenya, 1963–68	S	3.0				

Note.—For sources and notes, see Appendix B

References

Bhagwati, Jagdish. *The Economics of Underdeveloped Countries.* New York McGraw-Hill, 1966.

Heilbroner, Robert L. *The Great Ascent.* New York: Harper & Row, 1963.

Organization for Economic Cooperation and Development. *National Accounts of Less Developed Countries, 1959–1968.* Paris: Org. Econ. Cooperation and Development, June 1970.

Stebbins, Richard P., and Amoia, Alba, eds. *Political Handbook and Atlas of the World, 1970.* Council on Foreign Relations. New York· Simon & Schuster, 1970.

United Nations. *Yearbook of National Accounts Statistics, 1969* New York: United Nations, 1970.

——— *Demographic Yearbook, 1970* New York United Nations, 1971

Factor Proportions and Comparative Advantage in the Long Run: Comment

Alan V Deardorff

University of Michigan

Findlay (1970) examined the effects of growth on comparative advantage in an open three-sector growth model in which only the consumption goods were traded He simplified his analysis considerably by assuming factor prices to be completely determined by the world prices of traded goods It is the purpose of this note to show that this assumption is both more powerful and more restrictive than Findlay seems to have realized

In Section I we demonstrate the power of the assumption by showing that if all of its implications are realized, then certain results that Findlay thought ambiguous are actually determinate Then, in Section II, we argue that the assumption is restrictive, because it requires that both traded goods be produced Since such incomplete specialization will not occur for all parameter values, we describe the effect of dropping the assumption on the behavior of Findlay's model and on his results

I. The Power of the Assumption

Findlay's economy produces three goods· two traded consumption goods, X and Y, the prices of which are given from outside, and a nontraded investment good, Z, the price of which adjusts to equate supply and demand domestically Findlay argued that the fixed prices of X and Y would serve to determine the prices of the two factors, capital and labor, which would then determine the input coefficients in each sector [1] Thus, he was able

---

This paper was written with financial assistance provided by NSF grant GS-3073 to support research in international economics at the University of Michigan I am indebted to Robert M Stern for his comments on an earlier version of the paper.

[1] Determination of factor prices by commodity prices requires all of the assumptions of the factor-price equalization theorem Findlay explicitly assumed perfect competition, the absence of transport costs and tariffs, and the "strong factor-intensity assumption." He did not mention the additional requirement that the country actually produce both traded goods (incomplete specialization). We look further at this requirement in Section II.

to state the equations of the model in the following linear form:[2]

$$a_{11}\frac{X}{K} + a_{12}\frac{Y}{K} + a_{13}k = \lambda, \qquad (1')$$

$$a_{21}\frac{X}{K} + a_{22}\frac{Y}{K} + a_{23}k = 1, \qquad (2')$$

$$-s\frac{X}{K} - s\frac{Y}{K} + (1 - s)k = 0. \qquad (3')$$

Here, K is the capital stock, k is investment per unit of capital (equal to the growth rate of capital since depreciation is ignored), λ is the labor-capital ratio, s is the savings propensity, and a_{ij} are the fixed input coefficients of the factors, i $(1 = L, 2 = K)$, per unit of the outputs, j $(1 = X, 2 = Y, 3 = Z)$

Findlay solved this system, using Cramer's rule, to get several properties of the model The solution involved the determinant, Δ, of the coefficients on the left-hand side of the above equations. Findlay thought the sign of this determinant was ambiguous, except when the factor intensity of the investment good, Z, lay between the factor intensities of goods X and Y It was therefore only to this special case that he applied the model.

It turns out, however, that the sign of Δ is not ambiguous at all, but depends quite simply on the relative factor intensities of X and Y alone The reason is that when world prices determine factor prices, they do so by equating unit costs of each of the commodities to their prices [3] Thus the coefficients a_{ij} are related as follows.[4]

$$wa_{1j} + ra_{2j} = 1, \qquad (j = 1, 2, 3), \qquad (i)$$

where w and r are the fixed wage of labor and rental of capital, respectively It is possible to use equations (i) to simplify Findlay's expression for Δ to get

$$\Delta = (a_{22} - a_{21})/w \qquad (ii$$

Thus, Δ is positive (negative) if Y is relatively capital (labor) intensiv compared with X.

Equation (ii) can be derived by solving (i) for a_{1j} and substitutin directly into Findlay's definition of Δ, but this mechanical exercise is n really necessary. Instead we use (i) to transform equations (1′)–(3′) into an equivalent system that is more easily solved. Comparing his equations with (i), we see that if we multiply (1′) by w and (2′) by r and add, the resulting equation will have unitary coefficients on the left-hand side Thus we form the following new system of equations

---

[2] Those equations which appeared in Findlay's article will be referred to in this n by the same arabic numerals that he used. We will number our own equations w small roman numerals to distinguish them from his.

[3] Findlay mentioned this property early in his paper and even illustrated it in fig. 1 when he showed all three unit isoquants with a common tangent. It is somew surprising therefore that he did not make further use of it

[4] Findlay defined his commodity units to make each commodity price unity, giving us a 1 on the right-hand side of (i).

$$\frac{X}{K} + \frac{Y}{K} + k = w\lambda + r, \tag{iii}$$

$$a_{21}\frac{X}{K} + a_{22}\frac{Y}{K} + a_{23}k = 1, \tag{iv}$$

$$k = s(w\lambda + r) \tag{v}$$

Equation (iii) is derived from (1′) and (2′) as just described, equation (iv) is identical with (2′), and equation (v) is the sum of (iii), multiplied by s, and of (3′) The advantage of this formulation over Findlay's should be clear · it contains only the capital-input coefficients, the ordering of which is easily interpreted in terms of factor intensities

Equation (v) tells us immediately the solution for the growth rate of capital, k, which compares favorably, in terms of simplicity, with Findlay's solution in his equation (4) Indeed, if we differentiate (v) with respect to λ we find

$$\frac{dk}{d\lambda} = sw, \tag{vi}$$

which is unambiguously positive regardless of relative factor intensities. Thus, the ambiguity of this derivative which Findlay noted is resolved, and steady states are stable regardless of factor intensities.

Incidentally, if we equate Findlay's equation (5) with our equation (vi) and solve for Δ, we get immediately the relationship we noted in our equation (ii) Since Δ appears in many of Findlay's results, it is possible to use our (ii) throughout his paper to simplify and remove ambiguities Alternatively, one could start from the beginning and rederive all of Findlay's results from the system of equations (iii)–(v). Either way, as the reader can verify, the effect is to get most of Findlay's results in more manageable form with fewer ambiguous qualitative implications.[5]

---

[5] Notable examples are his eq (12) which becomes $d\lambda^*/ds = 1/ws$ and his eq. (13) which becomes $d\lambda^*/ds = -k/ws^2$, which are unambiguously positive and negative, respectively Findlay devotes a good deal of attention to the ratio, e, of X and Y. While our approach does not notably simplify the solution for this ratio, we are able to get easily interpreted solutions for X/K and Y/K separately. They behave exactly as one would expect from the Rybczynski theorem, however, so that it is unnecessary to reproduce them here.

II. The Restrictiveness of the Assumption

The given world prices of X and Y serve to determine unique factor prices only if both X and Y are actually produced domestically Since there is nothing in the system of equations (1′)–(3′) to guarantee nonnegative solutions for X/K and Y/K, the assumption will be valid for only a subset of the possible values of the exogenous variables [6] Outside this subset, Findlay's equations are of no use, and the model must be reexamined

How, then, will the model behave if only one traded good is produced? The answer to this question appears to be quite complicated, since the model must now determine factor prices, commodity prices, and factor proportions within the producing industries, and the equations for doing this will not be linear Fortunately, however, the problem has already been solved[illegible]ewhere. For as long as Findlay's economy produces only one of the tra[illegible] goods, say X, then the process of trading part of its production of X for imports of Y at fixed prices is completely separate from the processes of production and growth [7] Indeed, the economy behaves exactly like Uzawa's (1963) two-sector growth model of a closed economy Thus, the analysis of Findlay's model can be completed by simply tacking on two Uzawa models, one with X and the other with Y as the consumption good, to explain behavior in those regions where only one of these goods is produced [8]

While most of Findlay's conclusions extend easily to this more complete model,[9] one of his conclusions must be reversed Findlay observed, incorrectly as we have seen, that steady states will be unambiguously stable only if the factor intensity of the investment good is intermediate between those of the two consumption goods This would require, however, that the investment good be relatively capital intensive compared with the consumption good in one of the Uzawa models used to complete the model It was in precisely such a case that Uzawa and others found the possibility of unstable steady states in the closed two-sector model. Thus, what is really required in order for steady states to be unambiguously stable in the completed version of Findlay's model is that the investment good be relatively labor intensive compared with both X and Y

---

[6] It is possible to identify this subset by setting first X and then Y equal to zero i eqq (iii)–(v) and solving the resulting equations, with λ endogenous, for the bounds on outside of which only one traded good is produced These bounds, it turns out, both l between the extreme factor-intensity ratios of the three goods, within which factor pric would be exogenously determined if all goods were traded This suggests, in an admitted crude way, that the likelihood of Findlay's assumption being valid is smaller than t likelihood of factor-price equalization in an otherwise equivalent model with all th goods traded.

[7] We assume here that the savings propensity is independent of the price of the inv ment good relative to the consumption goods

[8] Or where Findlay's equations yield negative production of one of the goods

[9] His comparative advantage results require a trivial modification, merely noting the ratio of X and Y is now either zero or infinite, depending upon which good is produ

References

Findlay, Ronald. "Factor Proportions and Comparative Advantage in the Long Run." *J.P.E.* 78 (1970): 27–36.

Uzawa, Hirofumi. "On a Two-Sector Model of Economic Growth. II" *Rev Econ. Studies* 30 (1963) 105–18

The Value of Changes in the Probability of Death or Injury

Michael Jones-Lee

University of York

One of the more puzzling questions encountered in public investment decision making (especially in the health and transport fields) concerns the value placed on improvements (or reductions) in the safety of human life.

Schelling (1968) appears to have been the first to suggest that any potential change in the fatal (or nonfatal) accident rate may be regarded as a (typically small) change in the probability of death (or nonfatal accident) for any particular individual Mishan (1971) takes a similar view of the problem and argues that, if alternative resource allocations are to be compared solely on the basis of "potential Pareto improvements," then public investment projects which increase the safety of human life will be recommended only if the total of sums which individuals would be prepared to forfeit to effect such probability changes (and the other desirable consequences of the investment) exceeds the capital costs of the projects concerned

Given the manner in which Schelling and Mishan formalize the problem, the central question is therefore the nature of the Hicksian "compensating variations" in wealth[1] for various changes in probability of death or injury In fact, even if Mishan's potential Pareto improvement criterion is rejected as a basis for allocative decisions, information concerning these compensating variations will still be relevant for all but those public investment decisions which completely ignore private preferences

The primary purpose of this paper is to develop a qualitative analysis of compensating variations for changes in the probability of an individual's own[2] fatal accident. In addition, brief consideration will also be given to

---

An earlier version of this paper was presented at the International Economics Association Theory Workshop held in Bergen in July–August 1971. It therefore reflects the many helpful comments and suggestions of the workshop participants and organisers, especially Jacques Drèze and Jim Mirrlees. Naturally, only I am responsible for remaining error, obscurity, and triviality

[1] See Hicks (1943) for a discussion of compensating variations.

[2] Variations in the probability of an individual's own death (or injury) are what Mishan calls variations in "direct" risks. By contrast, "indirect" risks are associated with the death (or injury) of others.

the manner in which the analysis might be modified to deal with (*a*) nonfatal accidents and (*b*) the death or injury of others Throughout the discussion all probabilities are to be interpreted as being of an essentially subjective nature. Since individuals are hardly likely to view their own death or injury as the subject of repeated experiments, the preconditions for the existence of objective (relative frequency) probability measures would seem not to be fulfilled, and subjective probability therefore appears to be the relevant concept in the current context Finally, in the interest of simplicity, the analysis is initially restricted to single-period probability changes However, the majority of public investments which affect safety will do so over extended periods of time, so that ideally one would wish to develop results for the general multiperiod case This generalization of the basic analysis is considered at a later stage in the paper

1. The Assumptions

It will be assumed that an individual faces two[3] eventualities for each level of wealth, W, with which he commences the current period· either he may live through the current period, in which case he and his dependents will enjoy[4] wealth W, or he may die during[5] the current period, in which case his dependents will enjoy whatever proportion of W is bequeathable (the possibility that the individual's dependents may also enjoy the payoff from insurance contracts in the event of his death is considered in the Appendix) Such a situation may be described by saying that there exist two relevant states of the world—"life" and "death " In general, the individual will be taken to be concerned not only with the level of wealth, W, with which he commences the current period, but also with the state of the world that actually occurs, typically preferring "life" to "death" At the beginning of the period the individual is assumed to face a number of alternative decisions, each of which determines a level of wealth, W (which is not subject to any uncertainty), and probabilities of occurrence for each of the essentially uncertain states of the world While Hirshleifer (1965, 1966) was the first to discuss this type of decision problem, his analysis was rather informal, and a rigorous foundation for a theory of choice under such circumstances has only recently been provided by Luce and Krantz (1971). The latter have shown that, provided an individual's preference ordering over alternative (conditional)[6] decisions obeys a restricted set of fairly plausible axioms, then under the circumstances described above he will behave as an "expected utility maximizer" That is, the individual will behave as if each of his alternative decisions yields probabilities p and $(1 - p)$ for states of the world "death" and "life," respectively, and he then selects that decision which maximizes the mathematical expectation of utility given by

---

[3] Since we are concerned only with variations in the probability of the individual's own death at this stage, there is no need for explicit recognition of the various outcomes of life and death for people other than the individual himself

[4] This is not intended to imply that all of W will be consumed during the current period but, rather, that the individual commences the current period with marketable assets worth W.

[5] Note that this formulation of the problem does not distinguish between the various times at which the individual may die during the current period (It is an essentially discrete-time formulation adopted for the sake of analytical simplicity.) This is clearly a deficiency, since the individual will presumably prefer to die later rather than earlier in the period. However, the deficiency disappears in the continuous-time formulation of the problem outlined in Section 3 below

$$E(U) = (1 - p)L(W) + pD(W), \tag{1}$$

where $L(W)$ is the utility associated with wealth W conditional upon occurrence of state of the world "life," while $D(W)$ is the utility associated with wealth W[7] conditional upon occurrence of state of the world "death"[8] The functions $L(W)$ and $D(W)$ are unique up to the same linear transformation In addition, it will be assumed that both $L(W)$ and $D(W)$ are continuous and at least twice-differentiable functions

Given that he will live through the current period, the individual will be taken to prefer more wealth to less and also to be financially risk averse, so that

$$\frac{dL(W)}{dW} > 0 \tag{2}$$

and

$$\frac{d^2L(W)}{dW^2} < 0 \tag{3}$$

Given that he will die during the current period, the individual will be assumed not to be misanthropic to his heirs and also *not* to wish them more rather than less financial risk, so that

$$\frac{dD(W)}{dW} \geq 0 \tag{4}$$

[6] Roughly speaking, Luce and Krantz's analysis differs from other axiomatic theories of choice under uncertainty such as Savage's (1954) in that the former is concerned with decisions which determine the probabilities with which states of the world occur, whereas in Savage's analysis decisions have no effect on the probabilities of occurrence of the various states Decisions in the Luce-Krantz analysis are referred to as "conditional" decisions and are clearly the relevant kind of decisions in the current context in which alternative decisions affect the probabilities of occurrence of the states "life" and "death"

[7] This is not meant to imply that the individual is able to bequeath all of W to his heirs but signifies merely that the bequeathable sum is related to current wealth Presumably, in the absence of opportunities for entering into life insurance contracts, only a fraction of W is bequeathable The availability of insurance on the other hand would seem to provide the possibility of augmenting bequeathable wealth at the expense of current wealth to such an extent that the former exceeded the latter Insurance is ignored in the current section but is discussed below in the Appendix.

[8] Both Drèze (1962) and Fromm (1968) employed the expected utility-maximizing assumption in considering the question of safety improvements However, in contrast to the purpose of the present paper, both authors were primarily concerned with the optimal private expenditure on safety improvement given a relationship between expenditure and risk reduction.

and

$$\frac{d^2D(W)}{dW^2} \leq 0. \tag{5}$$

Furthermore, it would seem reasonable to suppose that for any given level of wealth, W, the individual prefers state of the world "life" to state of the world "death," so that for all W,

$$L(W) > D(W) \tag{6}$$

Finally, we introduce a rather strong assumption which, although it is not crucial for the analysis which follows, does permit the unambiguous interpretation of certain results It will be assumed that the individual displays a more pronounced sensitivity to variations in either wealth or financial risk if he expects to live rather than die during the current period. (The most extreme example of this kind of relative sensitivity is the bachelor with no dependent relatives who is also unconcerned for the welfare of others This type of individual will be indifferent concerning the amount he can bequeath and will therefore be indifferent concerning the level of his wealth and its associated financial risk if he expects to die during the current period.) This assumption is taken to have the following formal implications for $L(W)$ and $D(W)$

$$\frac{dL(W)}{dW} > \frac{dD(W)}{dW} \tag{7}$$

and

$$\left|\frac{d^2L(W)}{dW^2}\right| > \left|\frac{d^2D(W)}{dW^2}\right| \tag{8}$$

In fact, the only circumstance under which this would not seem to be a plausible assumption is when only a very small part of current wealth is bequeathable. The standard of living of the individual's dependents would then be considerably reduced by his death, so that it might be that the individual would be more sensitive to variations in current wealth or financial risk if he expects to die during the current period. While this latter case is taken to be the exception rather than the rule, it must be admitted that the prevalence of life insurance contracts tends to cast some doubt on the matter. The question of insurance and its possible implications for the forms of $L(W)$ and $D(W)$ is therefore considered in the Appendix.

2. The Analysis

Suppose that initially the individual has wealth $\bar{W}$ (> 0) and faces a probability $\bar{p}$ ($0 < \bar{p} < 1$) of death during the current period. His expected utility is then given by

$$E(U) = (1 - \bar{p})L(\bar{W}) + \bar{p}D(\bar{W}), \tag{9}$$

where $L(\bar{W})$ and $D(\bar{W})$ denote, respectively, $L(W)$ and $D(W)$ evaluated at $W = \bar{W}$.

Now suppose that the individual is offered the opportunity to reduce the probability of his death during the current period from $\bar{p}$ to $p(< \bar{p})$. Given the restrictions on $L(W)$ and $D(W)$ implied by (2), (4), and (6), the individual will be prepared to forfeit a positive amount, V, in order to effect such a probability reduction. By definition, the maximum amount he will forfeit must be such as to leave him with the same level of expected utility as in the initial situation The quantity V will therefore be given by[9]

$$(1 - p)L(\bar{W} - V) + pD(\bar{W} - V) = (1 - \bar{p})L(\bar{W}) + \bar{p}D(\bar{W}) \tag{10}$$

Similarly, if the individual is to be induced to accept an increase in probability, then the minimum sum he will accept as compensation will increase his wealth to a level such that the augmented probability and wealth also satisfy equation (10)[10]

Clearly, V as defined by equation (10) is the Hicksian compensating variation in wealth for a change in probability from $\bar{p}$ to p. It is therefore a matter of some interest to establish the form of the functional relationship between V and p implied by (10). Since this will involve examination of a number of complex expressions, it will be convenient to introduce some notational simplification In what follows we shall therefore employ the following definitions.

$$L \equiv L(\bar{W} - V),$$

$$D \equiv D(\bar{W} - V),$$

$$L' \equiv \frac{\partial L(\bar{W} - V)}{\partial V},$$

$$D' \equiv \frac{\partial D(\bar{W} - V)}{\partial V},$$

$$L'' \equiv \frac{\partial^2 L(\bar{W} - V)}{\partial V^2},$$

$$D'' \equiv \frac{\partial^2 D(\bar{W} - V)}{\partial V^2},$$

[9] Provided that bankruptcy imposes no limitation on borrowing, it is possible for V to exceed $\bar{W}$ However, if bankruptcy does impose a borrowing limitation, then it is necessary to restrict V to be less than or equal to $\bar{W}$. Strictly speaking, (10) should then be written as a weak inequality with a complementary slackness condition for the restriction on V. For the sake of simplicity, the possibility of bankruptcy will be ignored in most of what follows, though it will be necessary to give it explicit consideration when dealing with the value of complete elimination of the risk of death.

[10] The quantity V will therefore be negative for increases in probability.

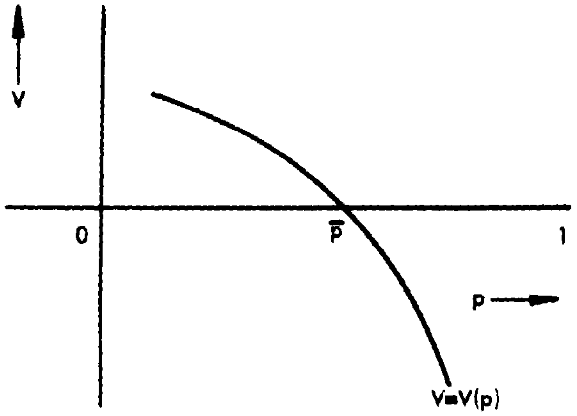


Fig 1

Finally, the functional relationship between V and p will be denoted by $V(p)$ Since p denotes a probability, it is clear that $V(p)$ is defined only on the closed interval $0 \leq p \leq 1$

Differentiating through (10) with respect to p and rearranging, we obtain

$$\frac{\partial V}{\partial p} = \frac{L - D}{(1 - p)L' + pD'} \tag{11}$$

It then follows that

$$\frac{\partial^2 V}{\partial p^2} = \frac{[(1 - p)L' + pD'](L' - D')(\partial V/\partial p) - (L - D)\{[(1 - p)L'' + pD''](\partial V/\partial p) + (D' - L')\}}{[(1 - p)L' + pD']^2} \tag{12}$$

But, from (2), (4), and (7),

$$L' < D' \leq 0, \tag{13}$$

while from (3), (5), and (8),

$$L'' < D'' \leq 0 \tag{14}$$

Hence, from (6), (11), and (13),

$$\partial V/\partial p < 0, \tag{15}$$

and from (6), (12), (13), (14), and (15),

$$\frac{\partial^2 V}{\partial p^2} < 0. \tag{16}$$

The graph of the relationship between V and p will therefore be of the general form depicted in figure 1. Note first that V is positive for values of

p less than $\bar{p}$, since the individual will be prepared to forfeit wealth to effect improvements in safety Conversely, V is negative for values of p greater than $\bar{p}$, since the individual will require compensation for deterioration in his safety. It should perhaps be emphasized that allocative decisions will frequently involve consideration of adverse as well as beneficial effects on the safety of life, so that the relationship between V and p for values of p exceeding $\bar{p}$ is of more than academic interest

Having established the general form of the relationship between V and p, let us now consider three aspects of this relationship in greater detail (*a*) the behavior of $V(p)$ as p approaches zero, (*b*) the behavior of $V(p)$ as p approaches unity, and (*c*) the derivative $\partial V/\partial p$ evaluated at $p = \bar{p}$

Clearly, (*a*) and (*b*) will involve individual attitudes toward extremes of safety and danger and will therefore be relevant only for those investment programs which have very pronounced effects on the safety of life However, $-(\partial V)/\partial p$ evaluated at $p = \bar{p}$ is the marginal value of a decrease[11] in risk from the initial risk level and is therefore directly relevant for all investment programs which have the effect of making life marginally safer (or riskier) [12] Since investments having relatively small effects on the safety of life are probably most common, it is clear that $\partial V/\partial p$ evaluated at $p = \bar{p}$ is of primary importance for policy purposes

a) The behavior of V(p) as p approaches zero

Suppose that $L(W)$ is such that

$$L^{-1}[(1 - \bar{p})\bar{L} + \bar{p}\bar{D}] > 0, \tag{17}$$

where $L^{-1}(U)$ is the inverse function[13] of $L(W)$, $\bar{L}$ denotes $L(W)$ evaluated at $W = \bar{W}$, and $\bar{D}$ denotes $D(W)$ evaluated at $W = \bar{W}$ The implication of (17) is that a positive value of W makes $L(W)$ equal to initial expected utility

---

[11] Note that $\partial V/\partial p$ is the marginal value of an *increase* in risk and is, therefore, negative By contrast, $-(\partial V)/\partial p$ is the marginal value of a decrease in risk and is consequently positive

[12] For a society of n individuals, the aggregate Hicksian compensating variation in wealth for changes in direct risk (in Mishan's sense, see n 2) due to some public investment program would then be approximated by

$$\sum_{i=1}^{n} \frac{\partial V_i}{\partial p_i} \delta p_i,$$

where δp_i is the (small) change in direct risk for the ith individual

[13] It is assumed throughout that this function is defined for all values of U that we consider; i e., it is assumed that if $L(W)$ is bounded below, then the greatest lower bound, γ, is such that $U > \gamma$ for all values of U considered, while if $L(W)$ is bounded above, the least upper bound, α, is assumed to be such that $U < \alpha$ for all values of U considered

In this case, setting $p = 0$ in (10) yields a value of V – say V_0 such that

$$V_0 < \bar{W}. \tag{18}$$

If (17) holds, then, the individual will not be prepared to bankrupt himself to make life completely safe during the current period,[14] and the graph of $V(p)$ will have an intercept $V_0(< \bar{W})$ on the V-axis.

If, by contrast, the inequality in (17) is reversed, then the individual will be prepared to bankrupt himself in order to reduce the probability of death in the current period from $\bar{p}$ to some value $\tilde{p}$ such that

$$0 < \tilde{p} < \bar{p} \tag{19}$$

In order to determine the behavior of $V(p)$ for values of p between $\tilde{p}$ and zero, it is necessary to establish whether the individual is able to borrow when bankrupt (i.e , when $W = 0$) If he is unable to do so, then, as noted earlier,[15] V will be subject to the constraint

$$V \leq \bar{W} \tag{20}$$

The function $V(p)$ will then be such that

$$0 \leq p \leq \tilde{p} \Rightarrow V(p) = \bar{W} \tag{21}$$

The graph of $V(p)$ will therefore become horizontal for values of $p \leq \tilde{p}$ Note that since (15) and (16) apply only when (10) holds as a strict equality, which will not be the case for values of $p \leq \tilde{p}$, there is no inconsistency among (15), (16), and (21)

If bankruptcy imposes no constraint on borrowing, then V will continue to increase as p is reduced below $\tilde{p}$. The graph of $V(p)$ will then have an intercept V_o' on the V-axis such that

$$V_o' > \bar{W} \tag{22}$$

Finally, suppose that

$$L^{-1}[(1 - \bar{p})\bar{L} + \bar{p}\bar{D}] = 0 \tag{23}$$

In this case the individual will not bankrupt himself for anything other than the complete elimination of the risk of death during the current period. The graph of $V(p)$ will therefore have an intercept V_o'' on the V-axis such that

$$V_o'' = \bar{W}. \tag{24}$$

b) The behavior of $V(p)$ *as* p *approaches unity*

It can be shown that if $L(W)$ is not bounded above, then it is possible to construct a variant of the Saint Petersburg game such that the individual will be prepared to pay an indefinitely large sum to participate in the game.[16] Since it is an observable fact that the majority of people will not pay indefinitely large stakes to participate in such games, it would seem most reasonable to regard $L(W)$ as bounded above If this is so, then by (6), $D(W)$ will also be bounded above. The least upper bound of $L(W)$ will be denoted by α and the least upper bound of $D(W)$ by β.[17]

---

[14] This is not intended to imply that it is in fact physically possible to make life completely safe during any period

[15] See n. 9 above.

Suppose that

$$\beta > (1 - \bar{p})\bar{L} + \bar{p}\bar{D}. \tag{25}$$

It follows that there exists a finite value of W which will make $D(W)$ equal to the initial level of expected utility There must therefore exist a finite value of V which will satisfy equation (10) when $p = 1$, so that there exists a finite sum sufficient to compensate the individual for the certainty of death during the current period. In this case, the graph of $V(p)$ will have a finite negative intercept on the line $p = 1$

Conversely, suppose that

$$\beta \leq (1 - \bar{p})\bar{L} + \bar{p}\bar{D} \tag{26}$$

In this case, no finite sum will be sufficient to compensate the individual for the certainty of death. Furthermore, it follows from (2) and (10) that

$$\lim_{p \to p^*} V = -\infty, \tag{27}$$

where

$$p^* = \frac{\alpha - (1 - \bar{p})\bar{L} - \bar{p}\bar{D}}{\alpha - \beta} \tag{28}$$

Note that if (26) holds as a strong inequality, then p^* will be strictly less than unity, so that there will exist a range of probabilities of death (namely, $p^* \leq p \leq 1$) for which it will be impossible to compensate the individual[18] The graph of $V(p)$ will then be asymptotic to the line $p = p^*$.

Figure 2 summarizes the various results that have been established concerning the behavior of $V(p)$ for extreme values of p.

c) *The derivative $\partial V/\partial p$ evaluated at $p = \bar{p}$*

When $p = \bar{p}$, it follows from (10) that $V = 0$. The derivative $\partial V/\partial p$ evaluated at $p = \bar{p}$, which will be denoted by $(\partial V/\partial p)_{\bar{p}}$, may therefore be obtained simply by setting $p = \bar{p}$ and $V = 0$ in (11), that is,

$$\left(\frac{\partial V}{\partial p}\right)_{\bar{p}} = \frac{\bar{L} - \bar{D}}{(1 - \bar{p})\bar{L}' + \bar{p}\bar{D}'}, \tag{29}$$

---

[16] See, e.g., Arrow (1965)

[17] The continuity of $L(W)$ and $D(W)$, together with the restrictions embodied in (6) and (7), implies that $\alpha > \beta$.

[18] Note that $V(p)$ is then defined only in the interval $0 \leq p < p^*$

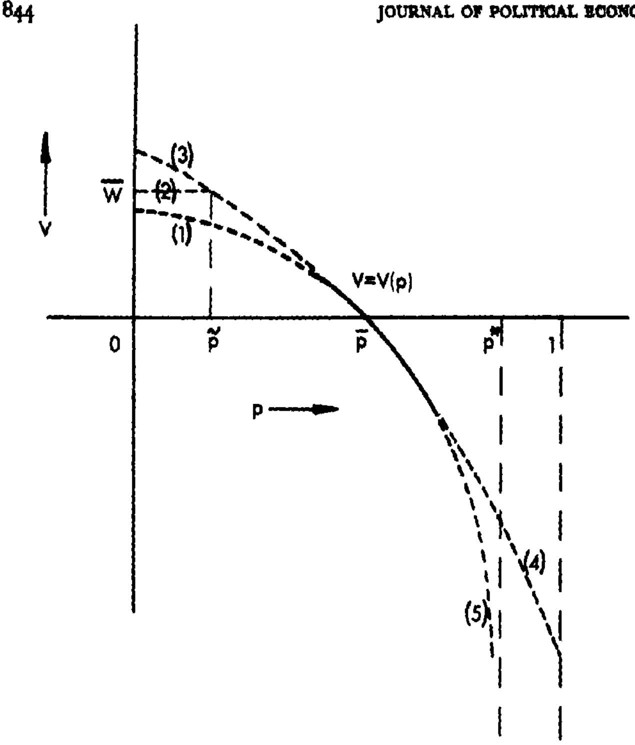


FIG 2.—(1), no bankruptcy at $p = 0$, (2), bankruptcy constraint for $0 \le p \le \bar p$, (3), no bankruptcy constraint, (4), possibility of compensating for death, (5), impossibility of compensating for $p^* \le p \le 1$

where $\bar L'$ denotes $\partial L(\bar W - V)/\partial V$ evaluated at $V = 0$ and $\bar D'$ denotes $\partial D(\bar W - V)/\partial V$ evaluated at $V = 0$

Now let us consider the manner in which the marginal value of a decrease in risk varies with initial risk, $\bar p$, and initial wealth, $\bar W$. Differentiating through (29) partially with respect to $\bar p$ gives

$$\frac{\partial(\partial V/\partial p)_{\bar p}}{\partial \bar p} = \frac{(\bar D - \bar L)(\bar D' - \bar L')}{[(1 - \bar p)\bar L' + \bar p\bar D']^2}, \tag{30}$$

while differentiating through (29) partially with respect to $\bar W$ gives

$$\frac{\partial(\partial V/\partial p)_{\bar p}}{\partial \bar W} = \frac{[(1 - \bar p)\bar L' + \bar p\bar D'](\bar D' - \bar L') + (\bar L - \bar D)[(1 - \bar p)\bar L'' + \bar p\bar D'']}{[(1 - \bar p)\bar L' + \bar p\bar D']^2}, \tag{31}$$

where $\bar{L}''$ denotes $\partial^2 L(\bar{W} - V)/\partial V^2$ evaluated at $V = 0$ and $\bar{D}''$ denotes $\partial^2 D(\bar{W} - V)/\partial V^2$ evaluated at $V = 0$ Thus, from (6), (13), and (30),

$$\frac{\partial(\partial V/\partial p)\bar{p}}{\partial \bar{p}} < 0, \tag{32}$$

while from (6), (13), (14), and (31),

$$\frac{\partial(\partial V/\partial p)\bar{p}}{\partial \bar{W}} < 0 \tag{33}$$

It will be recalled that the marginal value of a decrease in risk is given by $-(\partial V)/\partial p$ Thus, (32) and (33) indicate that the marginal value of a decrease in risk increases with both initial risk and initial wealth

3. Indirect Risks, Nonfatal Accidents, and Multiperiod Changes in Risk

This section outlines the manner in which the analysis might be modified to accommodate indirect risks (i e , outcomes of death or injury to others), nonfatal accidents, and changes in risk effective over extended periods of time

At the most general level, indirect risks and nonfatal accidents can be dealt with simply by extending the number of states of the world from "life" and "death" for the individual himself to the various mutually exclusive combinations of life, death, and injury for the individual and those he cares about The functions $L(W)$ and $D(W)$ of the earlier analysis are then supplemented by as many functions as there are additional mutually exclusive states of the world Clearly, the results of an analysis extended to cover indirect risks and nonfatal accidents will depend crucially on the relative magnitude of the various utility indices and on the signs and relative magnitudes of their first and second derivatives with respect to initial wealth It is because of the extensive restrictions required to generate determinate results in such cases that a purely a priori analysis is unlikely to prove particularly fruitful

However, the most significant single deficiency of the earlier analysis is that the changes in risk were taken to be effective only during the current period In order to expand the framework of the analysis to cope with changes in risk effective over extended time horizons, it would seem most satisfactory to deal with continuous rather than discrete time This may be done within the context of the Luce-Krantz expected-utility analysis by considering, as states of the world, "death at time t," where t takes on any value in some interval on the real line. One of the advantages of this procedure is that the potential temporal ambiguity associated with the outcome of death during, say, the tth period is completely eliminated, since it becomes possible to talk of death at a particular point of time

Suppose that the utility index associated with initial wealth W and death at time t is denoted by $U(W, t)$, the latter being a continuous and twice differentiable function of both arguments. If the (continuous) probability density function for the time of death is then written as $p(t)$, expected utility for initial wealth $\bar{W}$ and probability density function $\bar{p}(t)$ is given by

$$E(U) = \int_{\tau}^{T} U(\bar{W}, t)\bar{p}(t)\, dt, \tag{34}$$

where time τ is the present and time T is the latest date at which the individual conceives that he will still be alive The maximum amount which the individual would forfeit (or the minimum sum he would accept as compensation) for swapping the probability density function $\bar{p}(t)$ for some other density function $p(t)$ would then be given by V, where V is such that[19]

$$\int_{\tau}^{T} U(\bar{W} - V, t)p(t)\, dt = \int_{\tau}^{T} U(\bar{W}, t)\bar{p}(t)\, dt \tag{35}$$

Given appropriate restrictions on the form[20] of $U(W, t)$, it should then be possible to develop qualitative theorems concerning the value of different types of change in the probability density function, though it is not at all obvious which particular types of change should be considered in an a priori analysis of the "general case"

4. Concluding Comments

In the preceding sections we have succeeded in developing a set of interesting results concerning the value, in the form of Hicksian compensating variations in wealth, of current period changes in the risk of death. An attempt has also been made to outline the manner in which the analysis might be extended to accommodate (*a*) nonfatal accidents, (*b*) the risk of death and nonfatal injury to individuals other than the

---

[19] Note that eq. (10) above may be regarded as the special case of (35) obtained when $U(W, t)$ is the discontinuous function

$$U(W, t) = D(W)(\tau \leq t \leq t_1),$$
$$U(W, t) = L(W)(t_1 < t \leq T),$$

where t_1 is the time at which the current period ends

[20] E.g.,

$$\frac{\partial U}{\partial W} > 0, \quad \frac{\partial^2 U}{\partial W^2} < 0,$$

and

$$\frac{\partial U}{\partial t} > 0, \quad \frac{\partial^2 U}{\partial t^2} < 0.$$

However, the appropriate assumption concerning the sign of $\partial^2 U/\partial W \partial t$ does not seem at all clear a priori.

decision maker under consideration, and (*c*) changes in risk effective over an extended time horizon. However, further analysis of these three cases would require stronger specification of the utility functions (and the probability density function for time of death in case [*c*]) than seemed appropriate in the current context.

Of course, qualitative information is of only limited usefulness, and it is natural to wonder whether the analysis developed above might provide an effective basis for the derivation of quantitative information concerning the private values of changes in risk In fact, it seems that if it were possible to obtain some form of estimate of the utility functions $L(W)$ and $D(W)$, then derivation of the actual values of various single-period risk changes—and in particular the marginal value of changes in risk—would be a relatively straightforward matter Some variant of the experimental procedures developed by Mosteller and Nogee (1951) or Davidson, Suppes, and Siegel (1957) might well provide the basis for a procedure capable of generating estimates of the utility functions [21]

Appendix

Life Insurance Contracts[22]

Suppose that the individual enters into a contract with an insurance company whereby he pays a premium, R, in order that the beneficiary nominated in the contract should receive a sum, I (dependent on R), contingent on the individual's death [23] Clearly, this is a highly simplified version of the typical life insurance contract, but it does capture those features important for the present discussion.

The individual's optimal expenditure on insurance will be such as to maximize expected utility, $E(U)$, given by

$$E(U) = (1 - \bar{p})L(\bar{W} - R) + \bar{p}D(\bar{W} + I - R), \tag{36}$$

where

$$I = f(R) \tag{37}$$

and

$$R \geq 0. \tag{38}$$

For simplicity, the net insurance payoff will be denoted by N, so that

$$N = I - R. \tag{39}$$

If the individual insures in some nonzero amount and $f(R)$ is continuous and differentiable, then the optimal value of R will satisfy the necessary condition for a turning-point maximum of $E(U)$ given by

---

[21] For an example of a very imperfect attempt to obtain empirical information concerning the value of probability changes, see Jones-Lee (1969)

[22] My thoughts on this subject were greatly clarified by a discussion with Jacques Drèze.

[23] It is essential to appreciate that expenditure on the kind of life insurance under consideration is not intended to affect the safety of life in any way whatever. There may of course exist a "moral hazard" in the sense that the individual, once insured, will take a more reckless attitude toward his own safety. However, this possibility will be ignored in the analysis which follows.

$$(1 - p)\frac{\partial L(\overline{W} - R)}{\partial R} + p\frac{\partial D(W + N)}{\partial N}\frac{dN}{dR} = 0. \tag{40}$$

It will be assumed that the insurance payoff, I, is proportional to the premium, R, so that we may write

$$N = kR \ (k > 0). \tag{41}$$

Now insurance is typically categorized in the following manner[24]

$$\text{fair insurance} \quad k = \frac{1 - p}{p}, \tag{42}$$

$$\text{less than fair insurance} \quad k < \frac{1 - p}{p}, \tag{43}$$

$$\text{more than fair insurance} \quad k > \frac{1 - p}{p} \tag{44}$$

Thus, from (40), (41), (42), and (43) it follows that if the individual purchases fair or less than fair insurance, then

$$\frac{\partial D(\overline{W} + N)}{\partial N} \geq -\left[\frac{\partial L(\overline{W} - R)}{\partial R}\right] \tag{45}$$

But $\partial D(\overline{W} + N)/\partial N$ is simply $\partial D(W)/\partial W$ evaluated at $W = \overline{W} + N$, while $-[\partial L(\overline{W} - R)/\partial R]$ is equal to $\partial L(W)/\partial W$ evaluated at $W = \overline{W} - R$ Inequality (45) is therefore inconsistent with the restrictions on $L(W)$ and $D(W)$ given in (7) and (8) Consequently, in the case of an individual who purchases either fair or less than fair insurance of the type described, it would seem necessary to discard our earlier assumption concerning relative sensitivity to variations in wealth contingent on the life and death outcomes. The crucial question is therefore whether or not the observable prevalence of the "life insurance" type of contract (this being a generic term describing a wide variety of different contractual arrangements between insurers and insured) does in fact reveal a widespread willingness to undertake fair or less than fair insurance of the specific type described above There are at least two reasons for doubting that this is the case In the first place, insurance premiums are partially tax deductible in the United Kingdom and the United States Thus, even though insurance companies will typically offer less than fair insurance (in order to cover administrative costs and earn a return on capital), it may nonetheless be the case that after tax such insurance will be more than fair for individuals whose estates are sufficiently modest to escape death duties Moreover, the typical life insurance contract is for "endowment" insurance in which the face value of the policy accrues to the individual or his estate at the earlier date of the term of the policy or the death of the individual This type of insurance contract therefore provides "pure" life insurance jointly with an extended savings commitment.

For these reasons, there would seem to be very little justification for concluding that the popularity of life insurance implies a widespread willingness to enter into fair or less than fair versions of the type of insurance contract described by equations (37), (38), (39), and (41).

---

[24] See, for example, Friedman and Savage (1948)

References

Arrow, K. J. *Aspects of the Theory of Risk-Bearing*. Helsinki: Yrjo Jahnssonin Saatio, 1965.

Davidson, D.; Suppes, P ; and Siegel, S. *Decision-Making An Experimental Approach*. Stanford, Calif.: Stanford Univ, Press, 1957.

Drèze, J. "L'utilité sociale d'une vie humaine." *Rev. francaise recherche operationelle* 6, no. 23 (1962). 93–118

Friedman, M , and Savage, L J "The Utility Analysis of Choices Involving Risk." *J P E* 56 (August 1948): 279–304.

Fromm, G "Comment on Schelling." In *Problems in Public Expenditure Analysis*, edited by S B Chase Washington Brookings Inst , 1968

Hicks, J. R "The Four Consumers' Surpluses " *Rev Econ Studies* 11 (1943) 31–41

Hirshleifer, J "Investment Decisions under Uncertainty–Choice-Theoretic Approaches " *Q J E* 79 (November 1965): 509–36

———. "Investment Decision under Uncertainty Applications of the State Preference Approach " *Q J E.* 80 (May 1966) 252–77

Jones-Lee, M W. "Valuation of Reduction in Probability of Death by Road Accident " *J Transport Econ and Policy* 3 (January 1969): 37–47

Luce, R D , and Krantz, D H "Conditional Expected Utility " *Econometrica* 39 (March 1971) 253–71

Mishan, E J "Evaluation of Life and Limb A Theoretical Approach " *J.P.E.* 79 (July/August 1971) 687–705

Mosteller, F , and Nogee, P "An Experimental Measurement of Utility " *J P E* 59 (October 1951) 371–404

Savage, L J *The Foundations of Statistics* New York Wiley, 1954.

Schelling, T C "The Life You Save May Be Your Own " In *Problems in Public Expenditure Analysis*, edited by S. B Chase. Washington Brookings Inst., 1968.

A Reexamination of the Kinky Oligopoly Demand Curve

Walter J. Primeaux, Jr., and Mark R. Bomball

University of Mississippi

Over 30 years have elapsed since the kinked demand curve was introduced as a theoretical hypothesis to explain oligopoly pricing behavior.[1] There have been no empirical studies confirming the pricing behavior predicted by the hypothesis, and two separate studies by G. Stigler and J. Simon have rejected the theory,[2] yet the kinky demand curve persists as *the* theory of oligopoly for many economists. Obviously there must be some explanation for the tenacity with which economists cling to the kinky demand curve. At least three possible explanations exist for the survival of the theory up to this time: (1) perhaps the profession has elected to ignore the previous empirical evidence rejecting the theory; (2) perhaps there are problems in the previous studies which make their results something less than conclusive, and (3) the profession (because of the second possibility) requires additional evidence prior to passing final judgment on the theory. Apparently the explanation is that there is a need to generate more empirical evidence and tests of the kinky demand curve theory; that is the main purpose of this paper.

Broadly, the results of this study fail to support the kinked demand curve hypothesis. The results of this analysis, therefore, are generally consistent with those of both Stigler and Simon in their earlier studies.

The authors are indebted to George J. Stigler, Almarin Phillips, Julian L. Simon, Robert J. Larner, Coldwell Daniel III, Robert H. Rasche, and an anonymous referee for their helpful comments and suggestions. Charles G. Moscarillo of the Federal Power Commission was very helpful in supplying data and information. A portion of the research for this paper was conducted while Primeaux was a visitor at the Economics Department of the University of Pennsylvania preparing another paper for the Brookings Institution. The first paper, "A Re-Examination of the Monopoly Market Structure for Electric Utilities," was presented at a Brookings conference on Antitrust and Regulated Industries, October 28, 1971; publication is forthcoming. Throughout this paper, that project is referred to as the original study.

[1] The theory is generally credited to Sweezy (1939). Similar pricing behavior is also discussed by Hall and Hitch (1939).

[2] Stigler 1947 and 1968, Simon 1969. The Stigler pages cited here refer to the 1968 version.

Previous Studies

Stigler's landmark study examined price rigidity in industries which were monopolies and compared the results with price rigidity in oligopolistic industries. Stigler in testing the theory argued that "the kinky demand curve would prove to be an incorrect or unimportant construction if oligopoly prices were as flexible as monopoly and/or competitive prices."[3]

Although Stigler's investigation rejected the kinky demand curve hypothesis, there were several problems associated with his study which could have affected its results. First, there were several data problems generally acknowledged by Stigler in the article.[4] Second, Stigler's analysis did not involve very many monopoly producers for comparison with oligopolists, only two monopolies are included in his study.[5] Third, the firms examined by Stigler produced products which are heterogeneous between the monopoly and oligopoly market structures (for example, automobiles, steel, nickel, and aluminum, among others), to the extent that price elasticity of demand for a product affects pricing behavior, and to the extent that this elasticity varies among products, some of Stigler's counted price changes among and between industries could be the results of product differences, rather than differences in market structure.

Simon's study dealt with magazines and was also concerned with examining the number of price changes among firms to determine if there was, in fact, more price rigidity in oligopoly than in monopoly. Simon's study also rejected the kinked demand curve theory, but his analysis was similarly based on data which were characterized by certain weaknesses. First, as indicated by Simon, the status of any magazine as an oligopolist is doubtful, because the audiences of various magazines differ from one another so they are not perfect substitutes,[6] since various magazines constituted his universe, perhaps the outcome of the study was affected by this problem. Second, Simon was dealing with published advertising rates, there could have been haggling with major purchasers of advertising over the actual transaction price.[7]

Methodology

The procedures used in this study are very similar to those used by Stigler and Simon. The essential difference exists in the universe and sample used which overcomes virtually all of the data problems of the previous studies. The focus of attention is upon the relative number of price changes of a subset of firms in monopoly markets compared with another subset of firms which are oligopolists. Stigler, as quoted earlier, asserted that if oligopoly prices were as flexible as monopoly and/or competitive prices, the kinky demand curve would prove to be an incorrect or unimportant construction[8]

---

[3] Stigler, p. 221.

[4] Ibid., pp. 222–23. At the time his study was made, given his industries, these difficulties could not be avoided.

[5] Ibid., p. 224.

[6] Simon 1969, p. 972.

[7] Simon, however, indicates that he did not consider this a serious problem (ibid., p. 971).

Although frequently overlooked by economists, there are electric utility duopolies in the United States.[9] In the cities of interest, a municipally owned electric firm was in *direct* competition with a privately owned firm, and there was duplication of electric facilities The customer in these cities had a choice of being served by one firm or the other. In the Texas and Missouri cities included, for example, a customer could switch from one firm to the other at will In Portland, Oregon, on the other hand, new customers could take service from either company, however, customers could not switch from one electric firm to the other in that city This contrast is made so that the competitive conditions in these cities will not be confused with those existing in cities where territories are allocated and duplication of facilities does not exist Cities of this type are of no interest to this study

The methodology involved an examination of price changes in time-series data for each firm operating in cities with two electric utility firms. Such an examination would reveal oligopoly pricing behavior For purposes of comparison with the pricing behavior of the duopoly firms, another subset of firms was selected which was composed entirely of monopoly utility firms The monopoly subset consisted entirely of municipally owned firms which were virtually free of price regulation In the duopoly subset, a municipally owned firm competed with a privately owned firm, in these cases, the privately owned firms were subjected to some sort of price regulation at either the state or local level The municipally owned firms generally were free from price regulation The results of this study, as will be explained later, were not affected by these considerations

One may be inclined to question the comparison of municipally owned and privately owned electric utilities because of possible differences in profit incentives between firm types This tendency may raise two points concerning the procedure outlined above First, it may be asked whether the comparisons are valid, that is, it may be asked whether municipal firms behave like private firms since they are "nonprofit" institutions and thus not profit oriented as privately owned utilities Second (because of the first question), the question may be raised as to whether these nonprofit monopoly firms are a valid control with which to compare the behavior of the duopolies in the study. Although the common impression is that municipally owned firms are nonprofit, public-serving firms, their profit performance indicates that they are indeed profit oriented. For example, after discussing the level of profit for privately owned electric firms from 1950 to 1961, W. G. Shepherd explained that "despite the common impression, the *average* rates of return for publicly owned electricity suppliers (other than Federal projects) have been almost identical with these private profit rates."[10] Moreover, data from the Federal Power Commission also indicate that municipally owned electric utility firms seek profits. These data show that in 1968 net income amounted to 15.4 percent of sales for privately owned electric utilities and 22.3 percent for municipal electric utilities.[11] P. Mann found that municipally owned firms in his study earned a 7.14 percent rate of return on net investment while privately owned firms earned 6.74 percent.[12] M. Colberg has also discussed the profit orientation of municipally owned utility firms.[13] Thus it is obvious that municipally owned electric firms are not nonprofit institutions, and they are indeed motivated to seek profits. The fact that the average profit rates of municipally owned utilities are comparable with and sometimes exceed those of privately owned utilities seems to satisfactorily answer the two questions posed above, apparently municipally owned electric utilities' profit incentives are more like those of privately owned electric firms than is commonly supposed.

---

[8] Stigler, p. 221.

[9] Primeaux 1971.

As indicated above, the procedure used in this study eliminates virtually all of the data weaknesses of the previous studies. Oligopoly pricing behavior is observed where the products are homogeneous, for which there are no good substitutes for most uses; therefore, the influence of the differences in type of product on the number of price variations is eliminated. Moreover, the study examined an industry where rate schedules are not frequently varied, where deviations from printed rate schedules constitute illegal behavior for public utilities; there is no haggling concerning published prices as there could have been in previous studies. Also, since prices are not generally varied more than once during a given year, the possibility of distorted data is reduced. The data do not include years of depressed economic conditions which could abnormally reflect on firm pricing behavior. Another advantage of this approach is that there is a larger sample of monopolists than in Stigler's study. Furthermore, the duopolists in this study are clean-cut oligopolists, and there is no doubt concerning this matter as in Simon's study. Because of the advantages discussed above, after eliminating problems associated with previous studies, it should be possible to draw important conclusions concerning oligopoly pricing behavior by comparing the pricing variations of the duopoly subset with the monopoly subset.

---

[10] This refers to a return on capital (see Shepherd 1966, pp. 21–22).

[11] Information obtained from the Federal Power Commission, January 8, 1970.

[12] Mann 1970.

[13] Colberg 1955. This reference explains that electric profits were an important source of revenue for Florida cities and was a substitute for taxation.

The number of electric duopolies in this country is limited, therefore, it was necessary to include as many cities with that characteristic as possible. Even so, certain conditions discussed below made it necessary to exclude some electric duopolies from the study The monopoly subset was essentially that group included in the original study. Each firm in this subset was selected for the original study because of specific characteristics; the selection was on an entirely nonrandom basis.[14] The major criteria for their selection were as follows. First, to the extent possible, the monopoly firm should be from the same state as the competitive firms Second, the monopoly firm should be approximately the same size as the competitive firms; however, if no such firm existed in the relevant state a larger firm would be acceptable, and in no case would a monopoly firm be smaller than the competitive firms Whenever necessary to select a monopoly firm from another state, the substitutes would be chosen so that they would be situated in states geographically near to the state of the competing firm. Third, to the extent possible, types of power sources would be identical for the "matched" competitive and the monopoly firms.

The original study consisted of 23 duopolies and 24 monopolies The same sample was the original target of this study However, since several of the duopoly cities did not have competition during the entire 11 years included in this study, and several of the monopoly firms experienced competition at some time during this period, the decision was made to eliminate all firms for which the market structure had changed anytime during the years included. As a result of these eliminations, the final sample for this study consists of 17 cities with competition (since these are duopolies, this subset consists of 34 firms) and 22 municipally owned monopolies, which constitute the subset of matched firms. The firms finally included, and the cities in which they are located, are presented in the Appendix below.

Nature of the Tests

Two different methods were employed to test the kinked demand curve hypothesis. First, the number of price changes for the years included were computed for both the competitive and monopoly subsets; since the kinky demand curve predicts fewer changes in the duopoly case than in the monopoly case, the theory would not be upheld if the duopoly subset in this study displayed more price changes than the monopoly subset.

---

[14] Some firms were eliminated from this subset for various reasons which will be discussed below. In the earlier study, this subset was selected on the basis of important characteristics which are relatively unimportant to this study.

A second test consisted of counting the number of times price increases or decreases by one firm were followed by a competitor. The theory predicts that price decreases would be followed but price increases would not be followed

Number of Price Changes

The pricing information used was extracted from the FPC's *Typical Electric Bills* Each category or classification of price per kilowatt hour was treated as a separate and distinct product, and the number of price changes were calculated for each year In other words, 25 kwh of consumption was considered one product, 40 kwh another product, etc [15] The various categories or classifications used are presented in table 1 The procedure in this analysis follows very closely that employed by Simon [16] First, the number of price changes in each product category were computed This procedure yields the data presented in columns 1 and 3 of table 1 Then, the grand mean number of change years was computed by dividing the number of price changes in columns 1 and 3 by the number of firms in each type of market structure to arrive at an average which would make comparisons possible; otherwise, the effective number of price changes would be affected by differences in the sizes of the two subsets of market structures

The computations in table 1 reveal that there is significantly more price rigidity in the monopoly market structure than in the oligopoly market structure. These results reveal that monopoly electric utilities with no competition change prices less frequently than the firms in the duopoly market structure, these results are in contradiction with the kinky demand curve hypothesis and are consistent with those presented by both Stigler and Simon

Price Increases and Decreases

A fundamental proposition underlying the kinky demand curve is that rivals will follow price decreases but will not follow price increases This section examines this basic proposition of the theory. Using price information for the 1959–70 period (excluding 1963 for reasons indicated in n. 15), price changes were observed in the duopoly markets, and each rival's

---

[15] This procedure is consistent with that used by the Federal Power Commission (*Typical Electric Bills, 1969*, p 6) It was necessary to eliminate 1963 from the study because the FPC used uniform categories between 1959 and 1962, changed in 1963, and changed again in 1964. Between 1964 and 1970 the classifications were uniform but different from those used in 1959–62 Since unique categories existed in 1963, no comparisons were possible, thus the elimination.

[16] Simon 1969, pp. 972–73.

TABLE 1

PRESENTATION OF STATISTICAL DATA: NO OF PRICE CHANGES

1959–62			1964–70		
kwh	No of Price Changes (1)	Grand Mean No of Change Yrs (2)	kwh	No of Price Changes (3)	Grand Mean No of Change Yrs (4)
			Monopoly		
25	1	0 05	100	14	0 64
40	2	0 09	250	19	0 86
100	4	0 18	500	22	1 00
250	4	0 18	750	23	1 05
500	5	0 23	1,000	24	1 09
			Duopoly		
25	9	0 27	100	35	1 03
40	9	0 27	250	42	1 24
100	9	0 27	500	42	1 24
250	10	0 29	750	48	1 41
500	11	0 32	1,000	52	1 53

SOURCE —Tabulations for tables 1 and 2 made from data extracted from FPC's *Typical Electric Bills*, various years.

reaction to those changes were noted These data are presented in table 2 [17]

The data show that there were 42 instances during this period where price increases by one firm were not followed by its rival, this outcome

---

[17] Annual data are used instead of monthly data because the latter were unavailable A thorough search of the FPC's *National Electric Rate Book* as well as personal contact with the commission revealed many complex problems when attempting to reconcile data among firms on effective dates of rate changes In some cases firms changed their rate structure, discontinued discounts, changed the minimum bill charge, added fuel adjustment clauses, and changed water heater and space heater provisions Since rate schedules reported in the *National Electric Rate Book* were not standard among firms, evaluations concerning the real effect of such changes on rates were extremely difficult to assess accurately Therefore, it was decided that annual data from *Typical Electric Bills* would be more accurate for a comparison because they are presented in a standardized form through the use of a weighted average (see *Typical Electric Bills, 1968*, p 140). Obviously there are difficulties which arise when analyzing timing with something so crude as annual data Perhaps the greatest difficulty occurs from the fact that some firms may change prices more than once in a year and this would be obscured in annual data. Information from the FPC indicates that multiple price changes within a single year were uncommon for the years covered by the study. This fact may be verified by a review of various issues of the commission's *National Electric Rate Book*, therefore, this difficulty does not exist in this study Another timing difficulty emerges because of the inability to pinpoint the *exact* time of a price change by one firm compared with another firm The procedure of implementing lags to evaluate whether a price change was followed was used to help overcome this problem; this procedure is discussed later On balance, it is unlikely that the results of the study were biased one way or the other because annual data were used.

TABLE 2

PRESENTATION OF STATISTICAL DATA: PRICE INCREASES AND DECREASES

1. Price increases not followed (within 3 yrs)		42
2. Price decreases not followed (within 3 yrs)		42
3. Simultaneous price changes		48
Increases	20	
Opposite	10	
Decreases	18	
4 Delay in following price increases		18
1 yr	18	
5 Delay in following price decreases		15
1 yr	10	
2 yrs	5	
6 Delay in changing in the opposite direction		11
Municipal initiated a decrease 1-yr delay in increasing	2	
Private initiated a decrease 1-yr delay in increasing	4	
Private initiated a decrease 3-yr delay in increasing	5	

NOTE —A reconciliation of this table and table 1 is possible by the following procedure The first two lines of this table record independent action, not followed by the competing firm. Therefore, these changes total to 84 Lines 3–6 constitute reactions of firms to price changes initiated by the other duopoly firm Therefore, there were really *twice* the number of price changes indicated in lines 3–6 In summary, then, lines 1 and 2 account for 84 price changes and lines 3–6 account for 184, total, 268 changes. The dupooly section in table 1 (cols 1 and 3) accounts for 267 price changes. The difference of one price change between the two tables can be explained by the situation where one firm changed prices in one period, with no reaction from the competing firm, during the next period both firms changed prices According to the procedure established for counting price changes, this resulted in a double counting of a change in table 2 The results are not affected by this small difference

tended to support the kinked demand curve theory [18] During the same time, however, there were 42 instances where price decreases were not followed, and this outcome is in direct opposition to the kinky demand curve theory These aggregate results show that price decreases were not followed any more frequently than price increases, and these data, therefore, tend to refute the theory.

There were 48 simultaneous price changes during the 1959–70 period, *this means that in 48 instances both firms changed prices* Within this group of price changes, in 20 instances simultaneous price increases occurred, in 18 instances firms simultaneously decreased prices, and in 10 instances the direction of price change of one firm was not matched by the competing firm In other words, in the latter case, there were simultaneous changes but the direction of change of price for one firm was the opposite of that of competing firms. In the cases of both the simultaneous price increases and the simultaneous price changes in "mixed" directions, the patterns cannot be explained by the kinky demand curve Such pricing behavior does not conform to the patterns expected from the theory.

The point might be raised that regulatory restrictions on some firms may preclude free movement of prices in this type of business; therefore, it may be argued that the previous comparison does not really reflect behavior which is inconsistent with theory because instant price adjustments are not possible for the privately owned firms A lag was implemented to take the time constraint into consideration to eliminate that problem. Nevertheless, when one examines table 2, analysis of the delays in following price changes further refutes the theory After a 1-year lag, 18 price increases were followed This behavior, too, is contrary to the theory because the theory would not expect price increases to be followed at all Ten price decreases (fewer than the 18 price increases) were followed after 1 year, while five were followed only after 2 years The significant fact is that price increases were followed when the theory predicts that they would not be followed, price decreases were also followed after a time lag, however, only after a longer time and not even to the same degree

---

[18] In a situation where price leadership exists, the demand curve would not kink The original study explains that since a municipal firm faced a privately owned firm in all cities included in the sample, that it is doubtful that any collusion or anticompetitive behavior existed. A three year time period was arbitrarily selected to allow for any lag in price changes due to regulatory frictions.

These results reveal that price increases were more nearly simultaneous than were price decreases These results are consistent with those presented by Stigler and are the opposite of what would be predicted by the kinky oligopoly demand curve [19]

Although the point has not been previously discussed in detail, it is significant that the results fail to support the assumption of price rigidity in oligopoly behavior in spite of a bias which works against this outcome This argument is based on the notion that privately owned firms are, according to the conventional view, reluctant to ask regulators for price (rate) changes and open the way for rate hearings If this proposition is true, and the municipally owned firms are virtually free to manipulate prices at will, one would expect more price rigidity in the markets with regulated firms, even if only one firm is regulated Since the results show more price flexibility in the duopoly markets, the case against the kinked oligopoly demand curve emerges even stronger than if this bias did not exist

Conclusions

Like Stigler's and Simon's studies, the data from electric utility firms do not support the kinky oligopoly demand curve hypothesis The data show that oligopolists changed prices more often than monopolists and there was a marked tendency for oligopolists to follow a competitor's price increase, both of these outcomes are inconsistent with the theory

Empirical evidence concerning the kinky demand curve is still nonexistent, so perhaps the behavior observed in price wars, which do tend to support the assumptions of the theory, merely reflect temporary disequilibrium conditions in imperfectly competitive markets and may not apply to a general case of oligopoly conduct. In any case, after 30 years of waiting for proof of the theory, it seems appropriate for economists to concentrate on tests of this hypothesis and determine its appropriate niche in economic analysis

---

[19] Stigler, p. 221

Appendix

CITIES AND FIRMS ON WHICH TEST WAS CONDUCTED

Cities with Competition	Cities without Competition
Bessemer, Ala Alabama Power Co Bessemer Electric Service	Florence, Ala Florence Electricity Dept
Tarrant City, Ala Alabama Power Co Tarrant City Electrical Dept	Scottsboro, Ala Scottsboro Electric Power Bd.
Fort Wayne, Ind Indiana and Michigan Electric Co Ft. Wayne Electric Light and Power Works	Richmond, Ind Richmond Power & Light
Hagerstown, Md Potomac Edison Co Hagerstown Electric Light Plant	Bristol, Va. Bristol Va Utilities Bd
Bay City, Mich Consumers Power Co Bay City Electric Light Dept	Niles, Mich. Niles Board of Public Works
Dowagiac, Mich Michigan Gas & Electric Co Dowagiac Board of Public Works	Wyandotte, Mich Wyandotte Dept of Municipal Service
Traverse City, Mich Consumers Power Co Traverse City Light & Power Dept	Hillsdale, Mich Hillsdale Board of Public Works
Zeeland, Mich. Consumers Power Co. Zeeland Board of Public Works	Lansing, Mich Lansing Board of Water & Light
Kennet, Mo. Arkansas Missouri Power Co Kennet Board of Public Works	Sturgis, Mich City of Sturgis
Poplar Bluff, Mo. Missouri Utilities Co. Poplar Bluff Board of Public Works	Petoskey, Mich. Petoskey Municipal Light Plant

Appendix (Continued)

Cities with Competition	Cities without Competition
Trenton, Mo. Missouri Public Service Co. Trenton Municipal Utilities	Carthage, Mo. Carthage Water & Electric Plant
Cleveland, Ohio Cleveland Electric Illuminating Co Cleveland Dept of Public Utilities	Rolla, Mo Rolla Municipal Utilities
Columbus, Ohio Columbus & South Ohio Electric Co Columbus Division of Electricity	Columbia, Mo. Columbia Water & Light Dept.
Piqua, Ohio The Dayton Power & Light Co Piqua Municipal Light Dept	Marshall, Mo. Marshall Muncipal Utilities
Springfield, Ore Pacific Power & Light Co Springfield Utility Board	Springfield, Ill Springfield Water, Light & Power
Sioux Falls, S Dak Northern States Power Co of Minn Sioux Falls Light & Power Dept	Anderson, Ind Anderson Municipal Light & Power
Garland, Tex Texas Power & Light Co Garland Municipal Utilities	Logansport, Ind Logansport Electric Light Dept
	Eugene, Ore Eugene Water & Electric Board
	Greenwood, S C. Greenwood Commrs. of Public Works
	Watertown, S. Dak Watertown Municipal Utilities
	Springfield, Mo City of Springfield Utilities
	San Antonio, Tex San Antonio Public Service Board

References

Colberg, R. Marshall. "Utility Profits: A Substitute for Property Taxes." *Nat Tax J.* (October 1955): 382–87.

Federal Power Commission (FPC). *National Electric Rate Book.* Washington: Government Printing Office, various

———. *Typical Electric Bills.* Washington: Government Printing Office, various.

Hall, R L, and Hitch, C. J "Price Theory and Business Behavior." *Oxford Econ. Papers* 2 (May 1939) 12–45

Mann, Patrick "Publicly Owned Electric Utility Profits and Resource Allocation" *Land Econ.* 46, no 4 (November 1970): 478–84

Primeaux, W J., Jr "A Re-Examination of the Monopoly Market Structure for Electric Utilities" Paper presented at Brookings conference on "Antitrust and Regulated Industries," Washington, October 28, 1971

Shepherd, William G "Utility Growth and Profits under Regulation" In *Utility Regulation New Directions in Theory and Policy*, edited by W G Shepherd and T G Gies New York Random House, 1966

Simon, J L "A Further Test of the Kinky Oligopoly Demand Curve." *American Econ Rev* 59, no 5 (December 1969): 971–75

Stigler, G. J "The Kinky Oligopoly Demand Curve and Rigid Prices" *J P E* 55 (October 1947): 434–49 Reprinted in *The Organization of Industry* Homewood, Ill. Irwin, 1968

Sweezy, P M "Demand under Conditions of Oligopoly" *J P E.* 47 (August 1939) 568–73

Voucher Plans, Voting Models, and the Efficiency of Local School Finance

M L Eysenbach

University of Utah

Discontent with the present public school system is obvious, but ways to improve the system are not The search for new alternatives has led to serious consideration of the voucher plan proposed by Milton Friedman (Friedman 1962), and some experimental voucher schemes have been offered funding by the Office of Economic Opportunity Public interest in voucher plans seems likely to continue despite the Supreme Court's recent ruling against Ohio's tuition grants Economists particularly have been attracted by voucher plans that increase consumer choice and might improve production efficiency through greater competition among schools But discussion of the efficiency and equity effects of voucher plans has not been integrated into a political-economy model of voter behavior That task is undertaken in this paper The essential aspects of Friedman's voucher plan and a simple voting model are set out in Section I The implications of the voucher plan for voting behavior and hence the size of the school tax support are explored in Section II In the third section the model is applied to Robin Barlow's study of Michigan households (Barlow 1970) It is shown that a shift to a voucher plan may be neither efficient nor equitable

I

A simple voucher plan such as suggested by Friedman is used for the analysis in this paper Attendance laws are unchanged children must attend approved schools Such schools may be run by private or public bodies, may set their own fees, and may supplement their tuition incomes in any way they wish The only tax support any schools receive is voucher payments toward tuition fees The total school tax revenue (less administrative costs) is divided equally among all eligible children For each child, parents receive a voucher which may be used in payment of fees at any approved school If the school's fees exceed the voucher, parents pay the difference directly If the fees are less than the face value of the voucher, parents may not receive the difference in cash (and hence there is no incentive for a school to charge fees less than the voucher).

The political-economy model is straightforward: a citizen votes for a particular budget if his marginal cost in increased taxes is less than the marginal benefits he will receive from the choice of that budget compared to its alternative. The political process is viewed as a sequence of votes on increasingly large budgets: when American voters are asked to approve a total school levy, the implicit alternative is usually a smaller levy rather than no schools at all.[1] Marginal benefits per dollar of expenditure are assumed to decrease with increase quantities of schooling per pupil. It follows that if voters are ranked in terms of their net gains (marginal benefits less marginal costs) from a budget choice, the actual budget passed will be the largest one for which the net benefits are positive for the last voter required for a legal majority

II

A voter's benefits from school taxes may be divided into "internal" benefits associated with the schooling of one's own child, and "external" benefits associated with the schooling of other people's children The voucher plan will change voting behavior because it changes crucially the internal benefits which voters receive from their tax dollars Under the traditional system school taxes finance services in government schools, and the internal benefit is the amount a parent would be willing to pay for the schooling his child receives in these schools Under the voucher system school taxes finance vouchers which parents may use in place of cash toward payment of school fees The internal benefit is now the amount a parent would be willing to pay for this nontransferable, noncashable voucher (not the amount a parent would be willing to pay for schooling per se) Since a parent may buy schooling directly for cash, the maximum he would be willing to pay for a voucher is the face value of the voucher But the internal benefit need not be that large the internal benefit is less than the face value if parents would have preferred less schooling than that provided by the voucher Thus, the voucher follows the usual rules for a nontransferable gift-in-kind. the value of the gift to the recipient is equal to or less than the market price of the commodity received

A simple illustration may clarify this crucial difference between the two systems Suppose for some increase in taxes, school expenditures will be raised by $100 per pupil. Under the traditional system, this $100 per pupil is spent for government school improvements, improvements which some parents consider worth more than $100, but which others consider a frivolous waste. Under the voucher system, the $100 appears as an increase in the face value of the voucher. Parents who previously supplemented their vouchers by $100 or more now receive precisely a $100 benefit, that is, the saving in their out-of-pocket payments to the school of their choice. There is no reason for these parents to change the quality of the schooling they purchase for their children (aside from income effects if income redistribution results from the increased tax and increased voucher.) Other parents previously supplemented their vouchers by direct payments of less than $100 or not at all. The increased voucher now purchases some improvements in the quality of education received by their children, but we know that these improvements are not worth $100 to these parents since, prior to the tax and voucher increase, they did not choose to purchase the same improvements by direct, out-of-pocket payments

---

[1] Several recent school tax ballots in Washington have been explicitly of a sequential form, e.g., Levy A. *X* million dollars; Levy B: provided Levy A passes, an additional *Y* million dollars, Levy C· provided A and B pass, an additional *Z* million dollars.

In addition to the crucial change in internal benefits, the voucher plan might also change the external benefits from any level of school taxes. School taxes no longer determine the total expenditures on schooling since parents may supplement their vouchers An increase in taxes would result in no increase in total schooling bought, and hence would yield no increase in external benefits, whenever parents merely substitute voucher payments for out-of-pocket payment of school fees In general one would expect out-of-pocket payment of school fees mainly by wealthier parents, while for children of poorer parents, increases in tax level would result in increases in total schooling under the voucher plan as under the traditional system It is poorer children who figure in most arguments for the external benefits of public schooling· for the subsequent analysis in this paper, I assume that the external benefits of a given level of school taxes are unchanged by the voucher plan in both average level and in distribution by income class of voters [2]

To summarize, under the traditional system the internal benefit from school taxes is the value the parents place on the education of their child. Under the voucher system the internal benefit is the lesser of two figures: (1) the value parents place on the education of their child, or (2) the face (money) value of the voucher Therefore, if external benefits are unchanged with the voucher plan, the total benefits from any budget are less than or equal to the total benefits of that same budget under the traditional system unless the quality of schooling per dollar of expenditure improves with a change in the method of financing [3] The relative productive efficiency of each type of system is a controversy I will not enter here, for the subsequent analysis I assume no change in the quality of education per dollar spent

---

[2] Greater parental choice might affect the benefits of compulsory education as perceived by others. Advocates of uniformity will perceive a decrease in external benefits, advocates of diversity, an increase.

[3] To understand this reduction in the internal benefits from the school tax budget, it may help to recall that the value of any particular good depends on what substitutes exist. The voucher plan introduces direct parental financing of improved schooling, a close substitute to tax-financed schooling.

The general rule for all voters, "vote yes if increased total (i.e , internal plus external) benefits exceed marginal taxes," results in some economically interesting subordinate decision rules for specific cases. All voters who choose to supplement their vouchers have marginal internal benefits exactly equal to the increase in the voucher they receive For these voters, the general decision rule can be restated as "vote yes if increased external benefits exceed marginal net taxes" where net taxes are the difference between taxes paid and the face value of the voucher, that is, the subsidy paid to, or received from, others.[4] When these voters pay above-average taxes, they vote yes on a budget if their positive marginal external benefits exceed their marginal net taxes When these voters pay below-average taxes they vote yes on a budget unless their marginal net subsidy is offset by increases in negative external benefits [5] This voting rule for the individual, a direct comparison of external benefits and net taxes or subsidies, reminds one of the traditional welfare economist's tax-subsidy prescription for efficient resource allocation when externalities are present, but of course an efficient individual-decision role does not necessarily lead to an efficient collective decision in a voting situation

Nor does that decision rule apply to all voters Taxpayers who do not supplement their vouchers have marginal internal benefits less than the increased face value of the voucher They would vote for a tax increase only if the marginal internal plus marginal external benefits exceeded their increased taxes, it would be necessary but not sufficient for marginal external benefits to exceed marginal net taxes For these taxpayers an increase in school budget means an increase in the schooling provided their children, as well as an increase in the schooling of others—under the voucher system as under the traditional system, a voting choice analogous to a tie-in sale In contrast, those who value the education of their own children highly can, with the voucher system, separate the choice of superior quality for their own children from the choice of minimum quality of education required for all children However, this change in the constraints governing the choice of schooling for some children, the achievement of "marginal conditions" in this one area of resource allocation, in no way implies that the chosen budget is Pareto efficient The voucher plan schools are firmly in the world of second best

---

[4] Net taxes in the traditional system are the difference between taxes paid and the per pupil cost of schooling received in government schools

[5] Negative external benefits might arise if, for example, some parents viewed increased education of others as a source of increased job competition.

III

The exact size of the budget passed under the voucher plan would depend on the configuration of benefits and taxes. In this section one particular configuration, that found in Robin Barlow's study of Michigan households, is analyzed Barlow's figures indicated that local school taxes rose with a household's income, but benefits from public schools rose even more rapidly. Hence, voters ranked in terms of increasing net gains were also ranked in terms of increasing income; the poorer voters were net losers and the richer voters were net gainers at the level of budget actually passed with the traditional school system.[6]

Figure 1 is a diagrammatic presentation of the Barlow situation with school taxes and benefits rising with income level. On the vertical axis are the marginal costs and benefits of increases in the tax budget, and on the horizontal axis are voters ranked by income class. Budgets *X* and *Y* are additions of $100 per pupil to successively larger total budgets, for example, moves from $400 to $500 per pupil and from $500 to $600 per pupil, respectively A single marginal tax line is shown since the marginal tax burden of each voter does not change as the total budget increases. For simplicity linear functions are shown so that the crucial median voter is also an average taxpayer. If the legal majority is other than 50 percent, or taxes asymmetrically distributed about the median, the crucial-voter and average-taxpayer positions would not coincide, and the voucher level would be above or below the level of taxes paid by the crucial voter Either of these conditions would shift the relative political weight of the group of "richer" voters, that is, those paying above-average taxes, but would not alter the basic exposition of difference between the voucher plan and the traditional-schools plan.

The traditional school budget with marginal internal plus external benefits depicted by budget *Y* in figure 1 would just pass with a 50 percent majority rule. All voters to the right of *B*, the 50 percent mark, vote yes, and those to the left vote no; budget *X* would pass easily with all voters to the right of *A* voting yes

Would this budget *Y* pass if the school district adopted the voucher plan? The answer depends on the pattern of internal and external benefits which would together yield Barlow's pattern of total benefits rising with a voter's income Further discussion requires that each type of benefit be a nonrandom function of a voter's income; for simplicity, linear functions will be used.

At one extreme, if schooling has no external benefits, the school tax budget will fall to zero. The richer voters, who pay above-average taxes,

---

[6] Unless unanimity is required to pass any budget. Complications introduced by varying family size will be ignored.

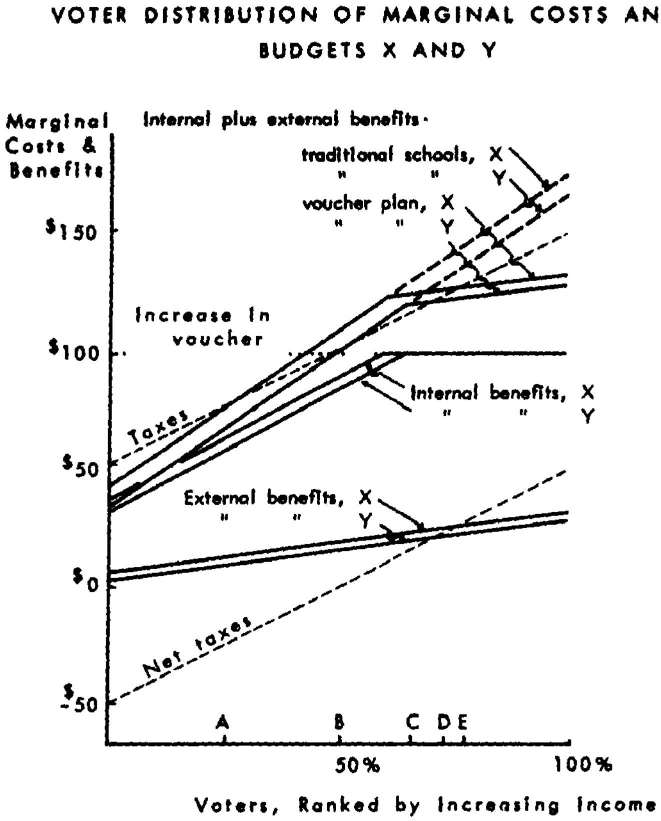


FIG 1

would vote no These richer voters pay taxes which exceed the face value of the voucher, since the voucher is equal only to the average tax; they always find it cheaper to pay school fees for their own children directly rather than through the tax system. With no external benefits, their net taxes necessarily exceed their external benefits at any positive levels of tax budget At the initial *Y* level of the school budget, the poorer voters, to the left of the 50 percent mark, also vote no They value their benefits from schooling at less than their taxes even though the face value of the voucher, and the cost of schooling, exceeds their taxes· the poor would prefer a lower level of both taxes and services At a lower budget level, marginal benefits exceed marginal taxes for some poorer voters, but the richer 50 percent of the electorate still votes no. Hence, no school taxes can be passed. (Obviously, if below-average taxpayers were numerous enough by themselves to constitute a legal majority, some budget could pass.)

At the other extreme the voucher budget is as large as the traditional budget. With Barlow's configuration of total benefits, the necessary and sufficient condition is that no richer voter's external benefits fall below his net taxes. If external benefits rise at least as rapidly as taxes, then external benefits exceed net taxes for all the richer voters at the budget level passed with traditional schools, these richer voters vote yes, and the same budget passes under either plan

Interesting voting changes take place when external benefits exceed net taxes for some richer voters as they do in figure 1 The internal benefits of the increased voucher of budget *Y* cease to rise with income class to the right of point *C* since the increased voucher merely substitutes for previous out-of-pocket payments by these parents To the right of *C* total benefits now rise only to the extent that external benefits rise with income class The marginal taxes of budget choice *Y* now exceed the marginal internal plus external benefits for all voters to the right of point *D* This dividing point is also characterized by net taxes exceeding external benefits Voters to the right of point *D* vote no under the voucher plan, and budget *Y* is defeated

A lower total budget, such as budget *X* in figure 1, would pass with the voucher plan At the lower level of total school budget we have higher marginal benefits for a $100 increase in school taxes per pupil, as shown, both the internal and external benefits are assumed to be higher for every income class At the lower budget level, the higher marginal schooling benefits exceed the school taxes of some poorer voters (those between points *A* and *B*) The higher external benefits exceed the net taxes of more of the richer voters (i e , those between points *D* and *E*). A new majority coalition is formed of the middle-income voters (between *A* and *E*), while the very poor and very rich still vote no.

We can now ask who might gain and who might lose from a change to this voucher plan The essential features of the change are· (1) the reduction, if any, in tax support, and (2) the freedom to purchase, without penalty, any quality of education above the voucher level [7]

One extreme case, the unchanged tax budget, is Pareto efficient the poor are no worse off, and the rich may now buy the more costly schooling they desire for their own children [8] At the other extreme, with no external benefits, there would be unanimous consent to the voucher plan with a slightly smaller tax budget, but as the tax budget falls the more prosperous poor become net losers. The zero budget actually passed with a majority rule is not Pareto efficient compared to the original budget.

---

[7] Under the traditional American system, if one wishes to purchase education in a nongovernment school, one must forfeit all internal benefits from the tax-supported school budget.

[8] Even this change would not be Pareto optimal if the traditional system permitted private schools: since all children are entitled to a voucher, the tax-supported expenditure per pupil must fall if the total budget is unchanged, since that total must now be spread over all students including former private school students.

The in-between case illustrated in figure 1 also is not Pareto efficient. Unambiguous gainers are the very poor and the very rich. The very poor, those to the left of point *A*, would prefer a still smaller budget than passes under the voucher system but their losses are reduced. The very rich, to the right of *E*, can now buy the better quality of education they want for their children They would be better off even if the budget level did not fall but in addition, with a smaller budget, the difference between the net taxes they pay and the smaller external benefits they receive is reduced

Unambiguous losers are the lower group of above-average tax payers (between points *B* and *C*) Under either system they would prefer a larger budget than budget *Y*, since even at the larger budget their marginal benefits exceed their marginal taxes; every reduction in budget means a net loss to them This loss is not offset by the freedom to buy a better quality of education for their own children because the original budget provided more schooling than they wanted for their own children

Three groups occupy an ambiguous position The richer voters between *C* and *D*, like the unambiguous losers, would prefer a larger school budget than is passed under their system, and hence every reduction in budget creates a net loss But members of this group, unlike the unambiguous losers, are willing to pay for better education for their own children and hence gain from the freedom to supplement under the voucher plan Their net gain or loss from the change to a voucher plan is uncertain The more prosperous poor (between points *A* and *B*) are the second ambiguous group They would prefer a smaller budget than passes under the traditional system—they vote no on budget choice *Y* under either plan—but would prefer a larger budget than passes under the voucher plan Whether they are better off with "overproduction" or "underproduction" cannot be determined from qualitative data alone. Similarly, those richer voters between *D* and *E* find the budget *X* with the voucher system too large and budget *Y* too small. Although they gain from the freedom to supplement, their net gain from the change of plan plus change in budget size is uncertain

Obviously, within the limitations of this model, there is no case for the voucher system on the basis of Pareto-efficiency. Nor is the equity case any clearer, for the only certain gainers are the extremes of very rich and very poor.

We have analyzed here the political economy of a voucher plan only within Barlow's configuration of taxes and total schooling benefits Alternative configurations of either benefits or taxes would yield different results The benefits of schooling may differ from Barlow's pattern, and the change in benefits as the results of a voucher plan will vary with the specifics of that plan, especially with provisions which limit parental freedom to supplement vouchers. Nevertheless, the results of this study show that the political economy of both simple and complex voucher plans requires more careful consideration by both their advocates and their opponents.

References

Barlow, Robin. "Efficiency Aspects of Local School Financing." *J.P.E* 78, no. 5 (September/October 1970): 1028–40.
Friedman, Milton. *Capitalism and Freedom.* Chicago Univ. Chicago Press, 1962.

Employer Costs and Discrimination: The Case of Baseball

James Gwartney and Charles Haworth

Florida State University

Economic theory implies that employers who discriminate are at a competitive disadvantage relative to firms that follow a less discriminatory policy. The low discriminator is able to reduce per unit cost because of his willingness to hire a disproportionate share of low-cost (relative to productivity) factors of production against which other firms discriminate. *Ceteris paribus*, the profits of the low discriminator will be greater, and his share of the market will tend to increase.[1] This theoretical proposition is difficult to test in the real world because of the lack of available indicators of discriminatory preferences and our inability to adjust for violations of *ceteris paribus* conditions between firms of an industry.

Analysis of desegregation in major league baseball provides a unique opportunity for researchers because in this industry it is possible to quantify the impact of a discriminatory policy on important revenue and cost variables.[2] In 1947, professional baseball teams were placed in an unusual position. A pool of readily available, proven baseball talent existed in the Negro leagues.[3] The contracts of most of these players could be purchased at a fraction of the cost that would be necessary to obtain comparable white players. Negro clubs were usually paid between $1,000 and $5,000 compensation for black players capable of transferring

The authors would like to thank David Beggs for his research assistance and two anonymous referees for their helpful comments on an earlier draft of the manuscript.

[1] See Becker (1957, chap. 3) for a complete statement of the theory. Becker concludes "In general, firms with discrimination coefficients less than the median discrimination coefficient are profitable and tend to expand relative to other firms" (p. 36).

[2] See Pascal and Rapping (1970) and Scully (1971) for evidence on the existence of employment discrimination in major league baseball. These two papers, utilizing primarily data from the sixties, focused on earnings and productivity differences between players according to color. Our paper focuses on the transition period (1947–59) and is a quantitative study on the impact of differing responses by individual firms to a desegregation of the industry. We are attempting to evaluate the payoffs (or costs) of being an innovator and/or a low discriminator in an industry that previously had been completely segregated.

[3] An indication of the talent available in the Negro leagues is the fact that six of the 12 "Rookies of the Year" selected during the 1947–53 period formerly played in the Negro leagues. During the 1950–59 period, eight of the 20 major league "Most Valuable Player" (MVP) winners had played in the Negro leagues. (See n. 12 below.)

directly to the major leagues. The highest price paid for a contract of a player from the Negro leagues was $15,000 (Peterson 1970). In contrast, the cost of player development for major league teams during the period 1945–50 was probably over $100,000 per major league player.[4]

Two inputs, white and/or black players, could be utilized in an effort to win games (an output variable), but the price of an equally productive black player was much cheaper. Once the color line was broken, *ceteris paribus*, firms had an incentive to substitute the less expensive blacks for the higher-priced white players.

There is considerable evidence that baseball executives were aware of the potential. As early as 1944, Bill Veeck tried to buy the sinking franchise of the Philadelphia Phillies and stock the club with stars from the Negro leagues. In his autobiography Veeck (1962) states: "With Satchel Paige, Roy Campanella, Luke Easter, Monte Irvin, and countless others in action and available, I had not the slightest doubt that in 1944, a war year, the Phillies would have leaped from seventh place to the pennant."

Veeck's plan to buy the Phillies was foiled. After a conference with the baseball commissioner, the owners sold the team to someone else for about half the price that Veeck was willing to pay. But 4 years later Veeck did gain control of the Cleveland Indian franchise. Paced by established players from the Negro leagues such as Paige, Easter, and Larry Doby, the Indians and Yankees dominated the American League in the late forties and early fifties.[5]

Other baseball men were also aware of the potential of black baseball players. John McGraw, in the early days of organized baseball, once tried to play a Negro, referring to him as an Indian (see Peterson 1970). Branch Rickey, the first baseball owner who was willing to break the color line, is reported to have felt that the employment of Negroes was necessary to restore his Brooklyn team that had been ravaged by war and old age.

---

[4] Data on player development costs for the New York Yankees and Chicago Cubs are available from the Celler Committee Report (U.S., Congress, House 1951). The Yankees estimated a development cost per player of $201,000, and the Cubs, $131,000. These figures include all of the losses of the farm system, which would be appropriate only if the whole purpose of the farm system was to develop major league players. Since there were some other advantages to farm systems such as retention of territorial rights, this is probably an overestimate. If no farm club losses are used (and some should be) the cost figure per player would be New York Yankees, $180,000, and Chicago Cubs, $90,000. During this period bonus payments were made almost exclusively to white players. The payments also are indicative of the cost of acquiring white player talent. A bonus is a payment to a free agent to sign a contract with a team. Between 1947 and 1958 over 100 white free agents were paid bonuses exceeding $25,000. The probability of any one of the bonus recipients becoming a major league player was one in three. These data are from Pascal and Rapping (1970) and various editions of the *Sporting News*.

[5] Between 1948 and 1956 the Indians won the American League pennant twice and finished second five times.

TABLE 1

Cumulative Number of Black Player Years, and the Won-Lost Record of Major League Baseball Teams 1952–56

Team*	Cumulative No. of "Black Player Years" 1947–56	1952–56	Games Won 1952–56 (%)	Ranking Won-Lost Record 1952–56
Brooklyn (N)	37	22	0.630	2
New York (N)	25	18	0.527	6
Cleveland (A)	20	12	0.619	3
Boston (N)	18	18	0.548	5
Chicago (A)	14	12	0.571	4
Chicago (N)	12	12	0.439	12
Cincinnati (N)	8	8	0.489	10
Philadelphia (A)	8	8	0.394	14
Pittsburg (N)	3	3	0.351	15
Saint Louis (A)	3	3	0.387	16
Saint Louis (N)	2	2	0.502	9
New York (A)	2	2	0.638	1
Washington (A)	1	1	0.432	13
Philadelphia (N)	0	0	0.509	8
Boston (A)	0	0	0.516	7
Detroit (A)	0	0	0.440	11
Average (top 8)	17.8	13.8	0.528	7.0
Average (lowest 8)	1.4	1.4	0.472	10.0
Average (top 5)	22.8	16.4	0.579	4.0

* National League teams are denoted by (N) and American League by (A).

In the 1940s, there were 16 major league baseball teams. The sport had sizable monopoly power, protected by a judicial interpretation providing baseball with exemption from antitrust prosecution. So there was little fear of a *new* innovative competitor, utilizing black players, entering and reducing the share of the market (and profits) of the existing firms. But, if an existing team was willing to employ blacks, it should have been able to increase its wins (output) per unit cost.

Just how large was the potential competitive advantage that could occur to a "low discriminator"? The ranking of teams according to the cumulative number of blacks on the team each year—"black player years"—clearly illustrates the contribution of black players to team success (see table 1).[6] By 1956 the Brooklyn Dodgers had accumulated 37 "black player years," 22 of them during the 1952–56 period. They were followed by New York (National League), Cleveland (American), and Boston (National). During the 1947–56 period, the eight teams with the greatest representation of blacks averaged 17.8 "cumulative black

---

[6] An additional "black player year" was assigned to a team for every additional black player on the team's annual official roster during the time period of the 25-player roster limitation of the league. The number of "cumulative black player years" is the number of blacks on the team's roster each year summed over a time period. Note that a player who is on the team's official roster for *X* years would contribute *X* times to the team's "cumulative black player years." The listing of Negro players from 1947 to 1959 by Bardolph (1959) was utilized to classify players on the basis of race.

TABLE 2

IMPORTANCE OF BLACK PLAYERS AS A DETERMINANT OF GAMES WON BY A TEAM DURING THE PERIOD IMMEDIATELY FOLLOWING THE FALL OF THE COLOR BARRIER

Equation No. and Dependent Variable	Intercept	X_1	X_2	X_3	R^2 (df)
(1) Games won per year (1950–55)	20.4	0.89		0.105	.42
		(4.23)		(5.84)	(93)
(2) Games won per year (1950–55)	16.6		3.75	0.111	.40
			(4.00)	(6.25)	(93)
(3) Games won per year (1950–59)	26.8	0.32		0.094	.32
		(3.63)		(7.39)	(157)
(4) Games won per year (1950–59)	22.5		2.05	0.101	.34
			(4.38)	(8.08)	(157)

NOTE.—X_1 = cumulative black player years, X_2 = number of black players during the *i*th year, X_3 = percentage of games won during the immediately preceeding 4-year period (1946–49).

player years." The comparable figure for the other eight teams was 1.4. The top eight teams in terms of "black player years" won 5.6 percent more of their games than did the eight teams with the least representation of black players during the 1950–56 period.

The evidence is even more impressive for the top five. The five teams that employed the most black players through 1956 comprised five of the top six teams according to won-lost percentage during the 1952–56 period. Only the New York Yankees, were able to attain this stature with relatively few black players. The five teams with the most black players won 58 percent of their games during the 1952–56 period. By contrast, the other 11 teams won only 46 percent.

Table 2, utilizing regression analysis, investigates the possibility that the teams with the most black players may have won more games during the early fifties because they started with better player personnel at the beginning of the period. The number of games won by a team during a season is the dependent variable. Equations (1) and (3) show that even after adjustment for initial team quality (X_3), the cumulative number of "black player years" had a highly significant positive impact on the number of games won during the *i*th year.

Equations (2) and (4) (table 2) consider the same two time periods, but the number of black players on the team during the *i*th year was substituted for the cumulative variable. Again, the evidence shows that the number of black players was a significant determinant of the games won by a team for both time periods. For the 1950–55 period, the results suggest that inclusion of an additional black player on a major league team, *on average*, resulted in an additional 3.75 wins per year, after adjustment for the initial team performance at the beginning of the period [7] For the longer 1950–59 period, each additional black player is estimated to have meant two additional wins per year for his team [8]

The importance of black players to team success is not surprising given their high productivity, on average, relative to whites Rosenblatt (1967) has previously shown that the batting averages of black players substantially exceeded that of whites [9] Table 3 presents an even more dramatic measure of productivity differences according to color The slugging averages of players according to color and position are presented for 1950 and 1955 [10] In each case, the slugging average of blacks, indicating both their ability to get on base and to hit for extra bases, subtantially exceeded that of other players even when grouped by position Both the batting and slugging average data highlight the potential of the available black playing talent during this era

Thus far we have emphasized the contribution of high-productivity and less expensive black players to team success as measured by games won However, if consumer discrimination against teams playing blacks was intense, team attendance, and therefore revenues, might suffer even while the team was winning Table 4 presents evidence that strongly contradict this hypothesis If consumer discrimination was strong, we

---

[7] To appreciate the importance of 3 75 games, it should be noted that in the period 1950–60, 3 75 games would have made the difference between winning a pennant and being runner up 30 percent of the time and the difference between second and third place over 50 percent of the time

[8] Annual games won was also regressed on "black player years" (X_2) and percentage of games won during the immediately preceeding 4-year period (X_3) for 1960–69 The number of black players per team remained significant at the 90 percent level, although the t-ratio (1 80) was considerably smaller for the sixties than for the fifties (eqq [2] and [4]) The size of the regression coefficient during the sixties was 0 77, less than one-half the comparable figure for the fifties Analysis of the partial coefficients of determination for the number of black players per year (X_2) suggests that the contribution of this variable to the explanation of the dependent variable was substantially smaller during the sixties The adjusted partial coefficients of determination for X_2 in eqq (2) and (4) were 131 and 098, compared to 007 for the same equation during the sixties

[9] The batting averages of Negroes and whites for 1953 and 1959, as indicated by Rosenblatt, were

Year	Negro Batting Average	White Batting Average	Difference
1953	288	257	031
1954	262	244	018
1955	271	251	020
1956	270	248	022
1957	264	252	012
1958	274	249	025
1959	265	246	019

Rosenblatt found that the differential continued to persist into the 1960s.

[10] A player's slugging average is his total bases on hits divided by his number of at-bats. *Ceterus paribus*, it will be higher for players who hit more home runs and other extra base hits.

TABLE 3

SLUGGING AVERAGE OF ALL PLAYERS AND OF BLACK PLAYERS BY POSITION FOR 1950 AND 1955

POSITION	SLUGGING AVERAGE			
	1950		1955	
	Blacks	All	Blacks	All
All .	.498	403	.440	.385
0F–1B .	492	426	453	.427
2B, 3B, SS .	481	372	421	358
Catchers .	550	387	.582	.380

would expect a negative relationship between home team attendance and the number of black players on the team Equations (1) and (4) (table 4) show that for both the 1950–55 and 1950–59 periods there was a highly significant positive relationship between team attendance and number of black players on a team, even after adjustment for the team's average attendance during the preceding 4-year base period (1946–49) On average, each additional black player on a team was associated with between 55,000 and 60,000 additional annual home-team admissions during the 1950s Undoubtedly, teams with black players had substantially higher attendance, at least partially because they won more games (see table 1) Equations (2) and (5) (table 4) estimate that team attendance was increased 12,000–15,000 by each additional game won during the 1950s Given our previous estimates that, on average during the fifties, each additional team "black player year" was associated with between two and four additional wins, the positive impact of black players on attendance is not surprising An additional black player meant additional wins, and the additional games won meant higher attendance

Equations (2) and (6) (table 4) estimate the impact of black players on team attendance, after adjustment for both base-period attendance and annual games won The number of black players is still positively associated with attendance, although the size of the black player coefficient, as expected, was reduced by the inclusion of the "games won" variable It was insignificant for the 1950–55 period but was positive and significant when the entire 1950–59 period was considered (eq [3]) Certainly these results indicate that black players not only helped win games but also brought customers to the ball park

In addition to their impact on team performance, two other factors may have contributed to the ability of black players to attract fans to the ball park The first might be referred to as the "superstar" phenomenon Outstanding players are thought to exert an additional impact on attendance because customers like to observe players of exceptional ability, even if they play for a "losing" team. The fact that blacks were overrepresented among superstars may, at least partially, explain their

TABLE 4

IMPACT OF GAMES WON AND BLACK PLAYERS ON ANNUAL TEAM HOME ATTENDANCE FOR 1950–55 AND 1950–59

INDEPENDENT VARIABLES	DEPENDENT VARIABLE—ANNUAL TEAM HOME ATTENDANCE					
	1950–1959			1950–55		
	(1)	(2)	(3)	(4)	(5)	(6)
Team black players (annual)	55,696		28,702	58,327		15,658
	(4.2)		(2.4)	(2.5)		(0.7)
Games won		15,193	13,772		11,973	11,369
		(8.2)	(7.2)		(5.7)	(5.0)
Average attendance, 1946–49	0.45	0.22	0.24	0.53	0.33	0.33
	(7.8)	(3.7)	(4.1)	(7.1)	(4.3)	(4.3)
Y-intercept	366,521	−404,195	−373,860	268,706	−336,808	−310,389
R^2	.33	.48	.49	.40	.52	.52
df	157	157	156	93	93	92

positive impact on attendance.[11] Second, black players may have attracted black customers who previously did not attend major league games. While no data on the racial composition of customers during the fifties are available, substantial attendance at Negro league games immediately after World War II and the rapid decline of the Negro leagues once the major league color bar was eliminated suggest that many black baseball fans switched their allegiance.[12]

Given their influence on home team attendance, what impact did black players have on team revenues during the fifties? Assuming that

---

[11] There is no precise definition of "superstar" available, but to illustrate this point two measures are suggested. At the end of each season an individual is voted by baseball writers as the outstanding performer in his league. During the period 1950–59, of the 20 most valuable players (MVP) selected, eight were black (40 percent). As an alternative measure of star status, one could use the All-Star Team Selection. Each year the outstanding players at each position, approximately 12 percent of all players, are chosen. The table below shows that blacks were dramatically overrepresented on these teams during the early 1950s.

Year	Blacks as % of All Players (1)	Blacks as % of All-Stars (2)	(2)/(1)
1950	1.7	8.0	4.7
1955	7.7	18.0	2.3
1960	10.5	20.0	1.9

[12] Many of the Negro league teams shared stadium occupancy with major league teams. Peterson (1970) reports that crowds of 12,000–15,000 were not uncommon in 1946 for Negro league baseball games. Over 45,000 fans were present for the East-West All-Star game during that year. By contrast, just 2 years later, attendance of less than 2,000 was common. This sharp reduction in attendance resulted in the demise of several teams. For example, desegregation of the Dodgers and Giants coincided with the demise of two New York entries (Newark Eagles and New York Black Yankees) in the Negro National League.

the net income per admission to a major league game was $1.00,[13] our attendance estimates suggest that the black-white differential revenue product was, on average, greater than $55,000 per black player. The potential financial gains available to low discriminators in this industry were large indeed For example, our estimates suggest that a low discriminator, willing to employ as few as four black players (among a 25-player team) could, on average, have increased his revenues by more than $200,000 annually, while simultaneously reducing the cost of acquiring major league players. The profit motive surely served to hasten the desegregation of baseball.

Conclusions

As is implied by economic theory, the empirical evidence shows that "low discriminators" (i e , teams willing to employ black players) obtained a competitive advantage in major league baseball relative to other teams They were able to win more games, acquire quality players at a lower cost, and increase annual revenue from admissions These gains offered an incentive for teams to desegregate

However, given the estimated contribution of black players to team success and financial gain, the results are not particularly encouraging as to the expected rate of decline in discriminatory barriers In the late 1940s and early 1950s, proven-quality baseball players were readily available from the Negro leagues Yet nearly 5 years after the fall of the color bar, less than one-half of the teams had desegregated [14] Perhaps major league baseball is unrepresentative of other industries because of the cartel power that the league has over the behavior of individual firms Additional evidence on other industries will be necessary before general conclusions can be drawn about time requirements necessary to change social attributes and reduce institutional barriers But in the case of baseball these factors did not change rapidly even when the opportunity for substantial financial gain was present

References

Bardolph, Richard *The Negro Vanguard* New York Rinehart, 1959.

Becker, Gary *The Economics of Discrimination* Chicago Univ Chicago Press, 1957

Pascal, A H , and Rapping, L. A. "Racial Discrimination in Organized Baseball " Xeroxed 1970.

---

[13] In 1950, after admission taxes, visitor share, and league share, the average income per admission was $1.05 (U.S , Congress, House 1951, p. 1619) This figure does not include any revenue from parking, concessions, radio-TV rights, and road receipts.

[14] One might argue that discriminatory barriers were reduced slowly in major league baseball because of the local monopoly possessed by each team. However, it should be recognized that despite the possession of monopoly power, the incentive to reduce costs and therefore obtain additional financial gain still remains.

Peterson, R. W. *Only the Ball Was White* Englewood Cliffs, N.J . Prentice-Hall, 1970.

Rosenblatt, A. "Negroes in Baseball· The Failure of Success." *Trans-Action* (September 1967), pp. 51–53.

Scully, Gerald W. "The Economics of Discrimination in Professional Sports The Case of Baseball." Paper prepared for a conference on Government and the Sports Business, December 6 and 7, 1971

U.S., Congress, House, Subcommittee on the Study of Monopoly Power *Organized Baseball* 82d Cong , 1st sess , 1951, ser. 1, pt 6

Veeck, Bill *Veeck as in Wreck* New York Putnam, 1962

Friedman versus Tobin: A Comment

T. Norman Van Cott

West Georgia College

Gary Santoni

Kansas State Teachers College

In a recent issue of this *Journal* Milton Friedman (1972) replied to five critics of his work. This short comment is directed only to Friedman's reply to James Tobin. This particular exchange was primarily concerned with putting the debate between the monetarists and neo-Keynesians into an *IS-LM* framework. Friedman, in disputing Tobin's contention that the argument reduces to one of determining the slope of the LM schedule, claims that the debate really centers around two other factors. First, "saving" and "investment" should be interpreted more broadly. Second, the monetary effect of any fiscal action will ultimately be predominant (i.e., an *LM* schedule continually shifting through time will "swamp" a once-and-for-all shift in the *IS* curve).

In this short comment we want to point out that Friedman makes a small mistake when discussing his first point in terms of the *IS-LM* paradigm. While the validity of his conclusion is not affected, manipulation of the model as Friedman suggests will always yield the right answer for the wrong reason. Were it not one of Friedman's major points in his exchange with Tobin, this comment would not be justified.

In referring to his first point Friedman (pp. 915–16) writes, "The monetarist view is that 'saving' and 'investment' have to be interpreted much more broadly than neo-Keynesians tend to interpret it, that the categories of spending affected by changes in interest rates are far broader than the business capital formation, housing construction, and inventory accumulation to which the neo-Keynesians tend to restrict 'investment.' Hence, even a fairly substantial tax increase will produce only a minor shift in the *IS* curve." Friedman goes on to point out that, regardless of the slope of the *LM*, a "minor shift" in the *IS* results in a minor change in income.

Friedman is incorrect in claiming that a broadened interpretation of saving and investment results in smaller shifts in the *IS* schedule in response to a change in taxes. This is not true for a change in taxes or any other change arising in the expenditure sector. The only effect of a broadened interpretation of saving and investment is to make the *IS* schedule "flatter." The leftward shift in the schedule in response to higher taxes always equals the tax increase times the relevant multiplier

Perhaps the simplest way to incorporate a broadened interpretation of saving and investment into the *IS-LM* paradigm is to introduce the interest rate as an argument in the consumption function Assume that consumption and the interest rate vary inversely as do neo-Keynesian investment and the rate of interest If we assume taxes are of the lump-sum variety, the equation for the *IS* curve reduces to

$$Y = \frac{a - bT + G_0 + g_0 + g_1 r + c_0 r}{1 - b}, \tag{1}$$

where Y is output, a is the constant term in the consumption function, b is the marginal propensity to consume, T is lump-sum taxes, G_0 is government spending, g_0 is the constant term in the investment function, and g_1 and c_0 show the responsiveness of investment and consumption, respectively, to interest rate changes Presumably, for the monetarists g_1 is greater in absolute value than it is for the neo-Keynesians and c_0 is equal to zero for the neo-Keynesians and less than zero for the monetarists

With reference to the interest rate axis, the slope of *IS* is

$$\frac{dY}{di} = \frac{g_1 + c_0}{1 - b} \tag{2}$$

That is, the more responsive investment and consumption are to the interest rate, the "flatter" the *IS* curve is

The shift in the *IS* curve in response to higher taxes is

$$dY = \frac{-b}{1 - b} dT \tag{3}$$

This is independent of g_1 and c_0

In terms of the *IS-LM* diagram (fig. 1) we have the following Assume IS_m is the *IS* schedule for the monetarists and IS_k is the neo-Keynesian schedule From an initial equilibrium at point A, a rise in taxes shifts the *IS* schedules leftward by the same amount, or, say, to IS_m' and IS_k' Note that the decline in income is smaller for the monetarists but *not* because of a smaller shift in the *IS* curve

One might object that Friedman is really referring to a vertical shift in the *IS* curve, in which case he is correct However, a reading of the section of his paper to which this comment is directed indicates that Friedman refers to shifting *IS* and *LM* schedules 10 times, with each shift described as a leftward or rightward shift

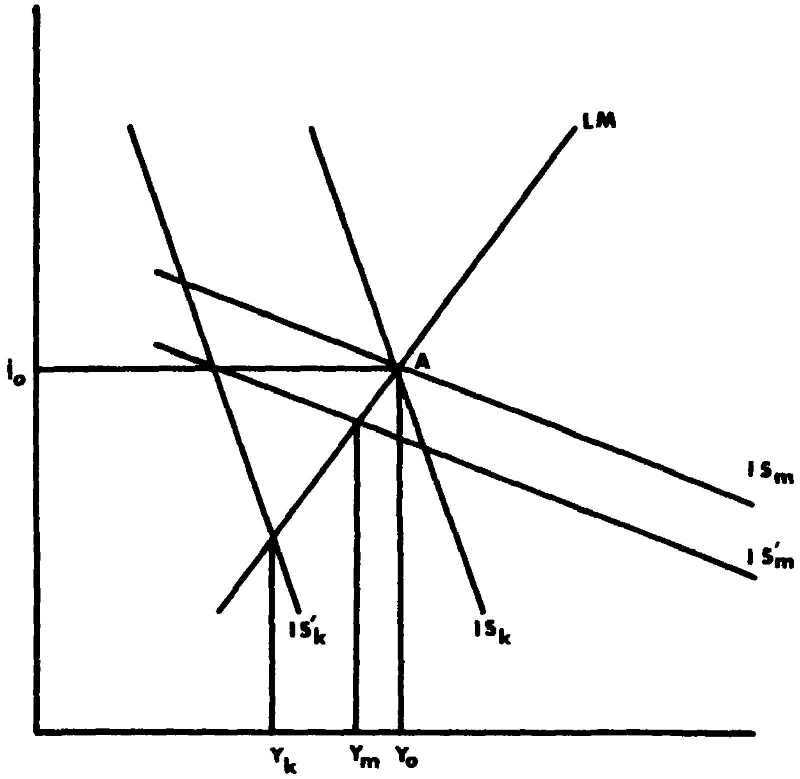


FIG 1

Reference

Friedman, Milton "Comments on the Critics" *J P E* 80 (September/October 1972): 906–50

Miscellany

The Economics of Brushing Teeth

The ever-growing literature on human capital has long recognized that the scope of the theory extends well beyond the traditional analysis of schooling and on-the-job training Migration, maintenance of health, crime and punishment, even marriage and suicide, are all decisions which can usefully be considered from the human capital point of view Yet economists have ignored the analysis of an important class of activities which can and should be brought within the purview of the theory A prime example of this class is brushing teeth [1]

The conventional analysis of toothbrushing has centered around two basic models The "bad taste in one's mouth" model is based on the notion that each person has a "taste for brushing," and the fact that brushing frequencies differ is "explained" by differences in tastes Since any pattern of human behavior can be rationalized by such implicit theorizing, this model is devoid of empirically testable predictions, and hence uninteresting

The "mother told me so" theory is based on differences in cultural upbringing Here it is argued, for example, that thrice-a-day brushers brush three times daily because their mothers forced them to do so as children Of course, this is hardly a complete explanation Like most psychological theories, it leaves open the question of why mothers should want their children to brush after every meal But it does at least have one testable implication· that individuals from higher social classes will brush more frequently

In these pages I describe a new model which is firmly grounded in economic theory and which generates a large number of empirically testable hypotheses. I then show that the predictions of the model are supported by the data.

Editor's note.—This paper derives from the Princeton oral tradition

I wish to thank my dentist for filling in some important gaps in the analysis, and my colleague, Michael Rothschild, for insightful kibitzing Support for this research is graciously solicited.

[1] The analysis to follow can also be applied to such important problems as combing hair, washing hands, and cutting fingernails, as I hope to show in a series of future papers.

The basic assumption is common to all human capital theory. that individuals seek to maximize their incomes. It follows immediately that each individual does whatever amount of toothbrushing will maximize his income The "mother told me so" model can be considered as a special case where the offspring only does as he or she is told, but the mother's decisions are governed by income maximization for the child Thus, offspring will behave *as if* they maximized income

An example will illustrate the usefulness of the model Consider the toothbrushing decisions of chefs and waiters working in the same establishments Since chefs generally come from higher socioeconomic strata, the "mother told me so" model predicts that they will brush more frequently than waiters In fact, it has been shown that the reverse is true (Barnard and Smith 1941) Of course, the human capital model predicts precisely this behavior On the benefits side, chefs are rarely seen by customers and work on straight salary Waiters, by contrast, are in constant touch with the public and rely on tips for most of their income Bad breath and/or yellow teeth could have deleterious effects on their earnings On the cost side, since wages for chefs are higher, the opportunity cost of brushing is correspondingly higher Thus, the theory predicts unambiguously that chefs will brush less It is instructive to compare this rather tight theoretical deduction with Barnard and Smith's glib attribution of the observed differences to the different hygiene standards in the birthplaces of the individuals (The chefs were born mostly in France, while the waiters were largely Brooklynites)

I. Review of the Literature

A substantial literature on dental hygienics exists It is ironic that economists are almost completely unaware of these studies, despite the fact that most economists brush their teeth

The best empirical study was conducted by a team of researchers at the University of Chicago Medical Center in 1967 They compared toothbrushing habits of a scientifically selected sample of 27 sets of twins who had appeared in Wrigley's chewing gum commercials with a random sample of 54 longshoremen The twins brushed their teeth an average of 3.17 times per day, while the longshoremen brushed only 0.76 times daily The difference was significant at the 1 percent level As noneconomists, the doctors advanced two possible explanations for this finding either twins had a higher "taste for brushing" than nontwins, or the Wrigley Company deliberately set out to hire people with clean teeth Further study, they concluded, would be needed to discriminate between these two hypotheses (Baker, Dooley, and Spock 1968). The human capital viewpoint makes the true explanation clear enough. Earnings of models depend strongly on the whiteness of their teeth On the other hand, no direct connection has ever been established between the income of longshoremen and the quality of their breath.

Another recent contribution was a survey of professors in a leading Eastern university. It was found that assistant professors brushed 2.14 times daily on average, while associate professors brushed only 1.89 times and full professors only 1.47 times daily. The author, a sociologist, mistakenly attributed this finding to the fact that the higher-ranking professors were older and that hygiene standards in America had advanced steadily over time (Persons 1971). To a human capital theorist, of course, this pattern is exactly what would be expected from the higher wages received in the higher professorial ranks, and from the fact that younger professors, looking for promotions, cannot afford to have bad breath.

II. A Theoretical Model of Toothbrushing

Let w be the wage rate of an individual, let J be an index of his job, and let B be the time spent brushing his teeth. With no loss of generality, I can reorder the jobs so that jobs with higher J are the jobs where clean teeth are more important. The assumed wage function is therefore

$$w = w(J, B), \qquad w_B \geq 0, \qquad w_{BJ} = w_{JB} \geq 0 \tag{1}$$

Since jobs have been reordered, there is no a priori presumption about the sign of w_J. It is also assumed that $w(\ )$ is continuous, twice differentiable, and semistrictly quasi-concave in the nonnegative orthant.

Each individual is assumed to maximize his income:

$$Y = w(J, B)(T - B) + P, \tag{2}$$

where T is the fixed amount of time per period available for working or brushing[2] and P is the (exogenously determined) amount of unearned income.[3] That is, each individual selects a value of B to maximize (2). The necessary condition for a maximum is[4]

$$w_B(J, B)(T - B) - w(J, B) = 0 \tag{3}$$

Several important implications follow from (3). First, since both w and w_B are presumptively positive, (3) implies that $T - B$ must be positive. In words, the theory predicts that no person will spend every waking hour brushing his teeth—an empirically testable proposition not derivable from either the "bad taste" or "mother told me" models.

---

[2] It is assumed, for simplicity, that these are the only possible uses of time. The model can easily be extended to accommodate an arbitrary number of uses of time, as is not shown in an appendix.

[3] A more general model would allow for the possibility that cleaner teeth can lead to a larger inheritance, that is, $P(B)$ with $P'(B) > 0$. For evidence of this, see "Toothpaste Heir Disinherited for Having Bad Breath," *Wall Street Journal*, April 1, 1972, p. 1.

[4] Since w is assumed semistrictly quasi-concave, this is also sufficient for a weak maximum.

Second, (3) can be rewritten

$$\frac{B}{T - B} = \frac{Bw_B}{w}. \tag{4}$$

In words, the ratio of brushing to nonbrushing time is equated to the partial elasticity of the wage with respect to brushing time. So individuals in jobs where wages are highly sensitive to brushing will devote more time to brushing than will others—as indicated in the verbal discussion Also, for any two jobs with equal w_B's but unequal w's, (3) implies that the higher-wage person will brush less due to his greater opportunity cost

Finally, consider the important case where (1) is linear in B (though possibly nonlinear in J)

$$w = \alpha(J) + \beta(J)B, \qquad \alpha \geq 0, \qquad \beta \geq 0 \tag{1'}$$

Substituting into (3) and solving yields

$$B = \frac{T}{2} - \frac{\alpha}{2\beta} \tag{5}$$

In jobs where brushing is immaterial to success, $\beta \rightarrow 0$, so (5) calls for a corner maximum with $B = 0$ Thus, we have a second strong prediction from the model· such persons will never brush At the other extreme, as the ratio α/β approaches zero, (5) implies $B \rightarrow T/2$ In words, individuals whose wages depend almost exclusively on the whiteness of their teeth (M C 's of television quiz shows are a good example) will spend approximately half their lives brushing Again, no sociological theory can generate predictions as strong as this

III. A Regression Model

The implications of the model can be put to an empirical test thanks to a recent cross-section study of American adults in the civilian labor force conducted by the Federal Brushing Institute. In its Survey of Brushing, the institute collected data on toothbrushing frequency and many socioeconomic characteristics of 17,684 adults in 1972 From these data, the following regression model was formulated

$$\begin{aligned} NBRUSH = {} & a_0 + a_1AGE + a_2WAGE + a_3NTEETH + a_4S \\ & + a_5EXP + a_6FDUM + a_7Y + u. \end{aligned} \tag{6}$$

The dependent variable is the number of times teeth were brushed during the year. AGE is included as a proxy for the number of years remaining before the individual's teeth fall out Viewing brushing as a human investment clearly implies that $a_1 < 0$. $WAGE$, of course, measures the opportunity cost of time; so $a_2 < 0$. $NTEETH$ is the number of teeth in the person's mouth. Since brushing time is nearly independent of the number of teeth brushed, having more teeth should certainly encourage more brushing. S and EXP are, respectively, years of schooling and work experience. They are included because this is a human capital model; although there are no a priori expectations about the signs of a_4 and a_5, both should have high t-ratios. $FDUM$ is a dummy for persons who live in an area with fluoridated water supply, included since there is some substitution in the production function for good teeth between brushing and fluoridating the water. Finally, Y is nonlabor income, which enables us to estimate the income effect on toothbrushing frequency.

Since I have argued above that $WAGE$ should depend on $NBRUSH$, equation (6) was estimated by the instrumental-variables technique. Denture wearers were included in the sample, but 189 people with no teeth at all were omitted from the analysis. The empirical results are reported below, with standard errors in parentheses.

$$\begin{aligned} NBRUSH = {} & \underset{(0.63)}{2.04} - \underset{(0.001)}{0.006}\, AGE - \underset{(0.001)}{0.096}\, WAGE + \underset{(0.009)}{0.054}\, NTEETH \\ & + \underset{(0.0002)}{0.0043}\, S - \underset{(0.0001)}{0.0022}\, EXP - \underset{(0.027)}{0.146}\, FDUM \\ & + \underset{(0.0002)}{0.0006}\, Y, \qquad R^2 = .79, \qquad SE = 0.056 \end{aligned}$$

By any standards the results are very good. The R^2 is very high for cross-section work, indicating that the data have been successfully mined. All the variables suggested by the theoretical model are highly significant and, wherever the theory implied a priori sign restrictions, they are satisfied.

In summary, the survey data strikingly confirm the predictions of the theoretical model of toothbrushing presented here. Of course, this is only one of many possible tests of the theory. But it does point out the usefulness of human capital concepts in understanding dental hygiene. Hopefully, these results will stimulate renewed interest in such questions on the part of economists.

ALAN S. BLINDER

Princeton University

References

Baker, M. D.; Dooley, C.; and Spock, B. "Brushing by Longshoremen and Twins: A Case Study." *Q. J. Orthodontics* 3 (1968): 377–462.

Barnard, C., and Smith, L. "Brushing Proclivities of Restaurant Employees in New York City." *Rev. Periodontics and Dentistics* 7 (1941): 1–2.

Persons, T. "Dental Hygiene and Age: A Sociological View." *J. Dental Soc.* 11 (1971): 1–243.

Proxies and Dummies

Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex

FRITZ MACHLUP

New York University

The Cambridge-Cambridge Controversy in the Theory of Capital; A View from New Haven: A Review Article[1]

Joseph E. Stiglitz

Yale University

Economics has long been plagued with controversies. the Keynesian-monetarist controversy, the investment function controversy (is the elasticity of substitution unity?), the liquidity preference–loanable funds controversy. Geoffrey Harcourt has extended his survey article from the *Journal of Economic Literature* (1969) into a book dealing with one of the latest of these so-called controversies, that between Cambridge, England, and Cambridge, Massachusetts, concerning capital theory The book is more balanced and more complete than the original survey article and includes some good pieces of exposition

The problems that I find with the book are basically problems I find with the Cambridge (U K) theory of which he is a partisan on one side—as, I suppose, I am on the other—and so, rather than focus on any errors and confusions which are peculiar to Harcourt, I prefer to focus on three of the major issues involved in the dispute and to suggest, in doing so, where Harcourt (and the Cambridge [U K] theorists) have gone astray

I. The Determination of Savings and the Rate of Interest

We all know that different individuals save different proportions of their income, depending on their age, number of children, income, the riskiness of their job, etc. Neoclassical economists have tended to emphasize the demographic characteristics (the life-cycle hypothesis), or, alternatively, when they wish a simplified exposition for students, they simply assume savings is a constant fraction of income The Cambridge economists have emphasized alternatively the role of institutions (corporations) in determining savings, and class behavior (whether the individual is a "worker"

I wish to acknowledge the helpful comments on an earlier draft of C. Bliss, J Mirrlees, R. Solow, D. Cass, and G Harcourt (although, not surprisingly, there wasn't unanimity on how the draft should be revised). Financial support from the National Science Foundation and from the Ford Foundation is gratefully acknowledged

[1] G C Harcourt, *Some Cambridge Controversies in the Theory of Capital* (New York Cambridge University Press, 1972).

or a pure "capitalist"). One would have thought that there would be an easy way to resolve such a controversy. which hypothesis provides a more accurate prediction of savings? Unfortunately, the hypotheses are not so easy to verify. For instance, direct evidence on savings of corporations may be misleading, for it is not unreasonable to assume that individuals can see, at least partially, through the corporate veil, and that corporate savings are a substitute for individual savings. (There appears to be some recent econometric support in favor of this hypothesis. Indeed, given our tax structure, it is to be expected that investment is largely financed by retained earnings.) Flow-of-funds analysis can provide relatively little information about the "real structure" of the economy

The question arises, however, Why the excitement over one specification of the savings behavior or another? The answer lies in the role that savings behavior plays in the Cambridge (U.K) theory in the determination of the rate of interest in long-run equilibrium In the simplest form, if savings are a given fraction s_π of profits, π, and all of wages are consumed, the rate of interest, r, is equal to the rate of growth, n, divided by s_π·

$$r = n/s_\pi \tag{1}$$

It is also true that with profit-maximizing competitive firms, in long-run equilibrium where all relative prices are constant, the rate of interest is equal to the own rate of return of every capital good[2] (the marginal productivity of every capital good in terms of itself):

$$r = MPK_{ii} \tag{2}$$

The controversy arises in the peculiar theory of causation. the Cambridge (U.K) economists argue that the rate of profit is *determined* by the rate of growth and s_π, and that this in turn *determines* the own marginal productivity of capital goods The reason for this view is that one can first solve equation (1) for r, without knowing anything about the production function or the factor supplies, and having solved for r, one can solve the second equation for the marginal productivity of each capital good and the capital goods–labor ratios

There are three objections to this view· first, there is a confusion between "causation" in a temporal sense and "causation" in the formal analysis of the structure of a set of equilibrium conditions It is only in the temporal sense that causation has any economic significance. At any moment, there is a given vector of capital goods and of labor. Under the extremely simplified models conventionally used, these endowments determine the marginal productivity of the different capital goods and the rate of interest. Given the savings behavior, this determines the change in the stocks of capital goods, eventually the economy converges to a state where the rate of interest is equal to the rate of growth divided by the savings propensity; still, at each moment, it is the "capital goods–labor ratios" which determine the rates of return on the different capital goods.

---

[2] I.e., the value of the marginal product of the capital good divided by its price. (I am assuming differentiability, so the marginal products are defined. See below.)

Second, in more general models, there is a simultaneity of the determination of the values of the variables of interest in the short- and long-run equilibrium.[3]

Third, the Cambridge (U.K.) approach cannot be extended to the determination of the relative factor prices if there are more than two factors; for example, if labor is not homogeneous, we have to have a theory which determines the relative prices of the different kinds of laborers. The marginal-productivity theory provides an explanation of all relative prices at the same time, the Cambridge theory has nothing to say except that, somehow, however the factor prices of the different kinds of labor are determined, the average wage must satisfy the "Cambridge" equation

II. Reswitching of Techniques

In comparing economies in balanced growth, one set of techniques may be chosen at a high and a low interest rate, and another set of techniques at an intervening interest rate This phenomenon—called "reswitching"—is the extension to a whole economy of the well-known phenomenon of multiple internal rates of return[4] All that this implies is that the weak

[3] For instance, if there are some savings out of wages, we replace eq (1) by $nK = I = s_\pi rK + s_w(Y - rK)$ (where K is the capital stock, I is investment, Y is output, s_w is the savings propensity out of wages), in which case eqq (1) and (2) must be solved simultaneously. Pasinetti (1962) has shown that if there are two classes, workers and capitalists (who do no work), then even though there is some savings out of wages, eq (1) still holds But the result is not robust if some capitalists save more than others, in the Pasinetti model r may be determined by the single capitalist with the highest savings rate, on the other hand, if workers save too much, then the pure capitalists disappear (in a relative sense), and eq (1) no longer obtains More generally, savings today will depend on expectations about the future, and these in turn will depend on the values of the relevant variables this period, which in turn depend on the level of savings.

[4] Assume we have a single consumption commodity, produced by labor and machines Machines last T years and require 3 years to construct Machine A requires a_1 units of labor the first year, a_2 the second, a_3 the third, machine B requires b_1, b_2, and b_3 in the first, second, and third years Which machine will be used depends simply on the sign of $(a_1 - b_1)\ (1 + r)^2 + (a_2 - b_2)\ (1 + r) + (a_3 - b_3)$ If $a_1 > b_1$, $a_2 < b_2$, $a_3 > b_3$, then there may be two values of r, say r_1^* and r_2, such that for $r \leq r_1^*$ and $r \geq r_2^*$ machine B is employed; for $r_1^* \leq r \leq r_2^*$, A is employed, and at r_1 and r_2^* the two machines are equally profitable In an earlier terminology, we would have said that this is an instance in which A is not an unambiguously more or less roundabout method of production than B; in present terminology, we say that A is not unambiguously more capital intensive than B. Note that consumption per man in a stationary state using A will be more or less than that in a stationary state using B as $a_1 + a_2 + a_3 \lesseqgtr b_1 + b_2 + b_3$. See Samuelson (1966) and the other articles in the *Q.J E* symposium on reswitching.

qualitative assumptions we conventionally make in economics—that is, convexity of the technology, with its implications of diminishing returns—do not have any strong implications for comparisons of economies in steady state. that is, reswitching establishes that the derivation of simple comparative dynamics propositions requires stronger assumptions than those required for the existence of competitive equilibrium and the derivation of qualitative properties concerning economies with given initial endowments. (In steady-state analysis, on the contrary, we let the initial capital stock be whatever it has to be to sustain the given steady state.) The conventional economic assumptions do enable us to make qualitative statements like "a decrease of a unit in the level of consumption today, c^0, will allow an increase in consumption tomorrow, c^1, of $\partial c^1/\partial c^0$, keeping consumption at all other dates constant, the one-period consumption rate of interest is equal to $\partial c^1/\partial c^0$ and decreases (more accurately, does not increase) with successive increases in consumption today." (If the technology is not differentiable, we must make the usual modifications involving left- and right-handed derivatives.)[5]

A comparison between the role of the "reswitching phenomenon" in capital theory and "Giffen's paradox" in consumer theory may be instructive. Competitive-equilibrium theory generally assumes convexity of preference sets, but to make qualitative propositions, even weak ones such as that demand curves are downward sloping, requires stronger assumptions. Giffen's paradox was important in the development of demand theory; the resolution of the change in demand into income and substitution effects is, most of us would agree, insightful. The reswitching phenomenon may eventually play a similar role in capital theory. But I doubt it, for two reasons.

1. Steady states are of limited interest in themselves, even the best of well-run economies never have a choice of steady states. They have a choice of consumption paths beginning with present initial conditions, conventional assumptions do allow us to make qualitative statements about such paths (economies may of course, eventually go to one steady state or another).

Moreover, it is easy to develop, using steady-state analysis, all manner of paradoxes.[6] For instance, the opening of free trade may actually lower steady-state consumption (this does not contradict the classical propositions concerning gains to trade).[7] One can show that of all the feasible steady states in a life-cycle model, the one which maximizes steady-state utility is not sustainable by a competitive equilibrium (without appropriate lump-sum transfers), and conversely (This does not contradict classical welfare propositions.)[8]

---

[5] That is, if $(\partial c^1/\partial c^0)^-$ and $(\partial c^1/\partial c^0)^+$ are the left- and right-hand derivatives, then the one-period rate of interest lies between $(\partial c^1/\partial c^0)^-$ and $(\partial c^1/\partial c^0)^+$.

[6] Elsewhere (Stiglitz 1973*b*) I have discussed the concept of *recurrence*, that in the process of economic growth one technique is used, then another, and then the economy returns to the earlier technique. Recurrence can occur in technologies in which reswitching does not occur, and need not occur in technologies in which reswitching does occur.

2. Though Giffen's paradox is irrelevant for most commodities, there are a few, perhaps potatoes and cheap wine, which have upward-sloping demand curves. There is thus some empirical basis for an interest in Giffen's paradox.[9] So far, no one has provided us with a corresponding empirical basis for an interest in reswitching. The reason, I suspect, is that the conditions under which it can be ruled out are very weak, for instance, reswitching cannot occur if, in only one industry in the economy, it is possible, by an increase in labor, either to reduce the input of some other factor, keeping other factors and output fixed, or to increase output, keeping all other factors fixed.[10] For reswitching to occur not only must there be essentially no substitutability in *every* industry, but also there must be large changes in the relative price for different capital goods, for it is the change in the relative price of capital goods which, in effect, leads one process to be more "capital intensive" at one interest rate (so the cost of using it rises more with an increase in the interest rate than the cost of using the other process) while it is less capital intensive than the other process at other interest rates. The question is, Are the kinds of changes of interest rates experienced in a growth process likely to be associated with such large changes in relative prices? Unfortunately, this is not a question which can really be answered, since most of the changes in relative prices which we observe have more to do with technical change, ignored in the discussion of reswitching, than with changes in interest rates, perhaps this simply testifies to the irrelevance of the reswitching controversy.

Thus, although reswitching has no implications for the validity of neoclassical distribution theory or for qualitative statements concerning the consequences of a given economy's increasing its level of consumption, it does mean that statements such as "the reason that one economy has a lower interest rate than another is that the former economy has a higher capital labor ratio" may be seriously questioned, for it is *possible* that a given economy have a lower interest rate than another and a higher interest rate than a third, but that the two comparison economies be identical in all physical respects. But by the same token statements like "sales of the commodity were high because the price was low" may be questioned, because it is possible for the higher sales to be associated with higher prices as well (It should be noted, however, that in comparing economies at the same time, which can trade capital goods with one another so that they have the same relative prices [approximately], these problems cannot arise)

---

[7] See Stiglitz (1970).

[8] Although the possibility, in the absence of money, of oversaving (rate of interest less than rate of growth) in a competitive equilibrium is an example of a competitive equilibrium which is not Pareto optimal. See, e.g., Samuelson (1958), Cass and Yaari (1967), and Diamond (1965).

[9] It is perhaps worth noting that the occurrence of Giffen's paradox depends on the level of aggregation chosen; wine is not likely to be Giffen, "cheap" wine is.

[10] Not only must there be this limited amount of substitutability, it must be "smooth substitutability"; i.e., the directional derivative must be defined. This is still much weaker than differentiability. For this and other "nonreswitching theorems," see Stiglitz (1973a).

Some of the confusion arises because in dynamic analysis, just as in static analysis, there is a distinction between macroeconomic and microeconomic models. Although we all recognize that there should be a close relationship between the two, the models which are appropriate for answering one class of questions may not be appropriate for answering another. Thus, to answer the kinds of questions conventionally raised in general equilibrium analysis, we never need to talk about an aggregate savings function; we need only impose fairly weak restrictions on each individual's preferences In macroeconomics, we want fairly precise qualitative (and indeed quantitative) answers to particular questions, such as the effect of a change in some policy This necessitates making stronger assumptions. The criticism of neoclassical growth theory is really an assertion that the restrictions embodied in neoclassical macroeconomic models do not necessarily follow from the microeconomic (disaggregative) models from which they should be derived. But the same criticism could be leveled against the static macroeconomic models.

Thus, to put the matter another way, if you accept (as most participants on both sides of the controversy do) the von Neumann (Sraffa-Samuelson-Solow, et al.) dual price–interest rate inequalities and the assumption that at any prices the cost-minimizing techniques are chosen, you have accepted the full content of the (microeconomic) neoclassical analysis of steady states. Reswitching says something about what that does or does not imply for comparisons of steady states, but no more.

III. Aggregate Capital

The first major attack of the Cambridge (U.K.) economists was not concerned with the saving hypothesis or the "reswitching phenomenon" (indeed, the latter was considered more of a curiosum, or a "perversity," than a serious economic problem) but with the existence of an aggregate capital stock. But like the reswitching phenomena, this is now recognized to be a red herring—and for very much the same reasons. Neoclassical distribution theory nowhere requires the use of aggregates, and to make qualitative statements of the kind discussed above requires only convexity of the technology.[11]

Wicksell (a neoclassicist!) was the first to discover the difficulties one encounters in making comparative dynamic statements involving the aggregate value of capital stock, consumption, and interest rates of economies in long-run equilibrium, he discovered, for instance, that changes in the consumption per unit of change in the value of the capital stock might not be equal to the rate of interest, because of changes in the relative price of capital goods[12] I have already argued that from a theoretical point of view, the analysis of steady states is not of much economic significance and may be misleading, and that the basic qualitative properties of economies out of steady state do not in any way depend on the ability to form aggregates From a practical point of view, economists are always dealing with aggregates one person's labor is aggregated with another, one piece of land is aggregated with another, one kind of steel is aggregated with another, even though they all have different properties The condition under which these aggregates can be formed, that is, under which the aggregates act as if they were homogeneous factors of production, are very restrictive, nonetheless, I believe that, under most circumstances and for most problems, the errors introduced as a consequence of aggregation of the kind involved in standard macroanalysis are not too important, nonetheless, we must always be on our guard for situations in which this is not true. The question is, Do the problems associated with the accumulation of capital in growth processes represent one area in which properly formulated aggregates (e g , using chain indices) are likely to lead to serious error? This, I suggest, remains a moot question

---

[11] In an attempt to deemphasize the role of aggregate capital, Solow (see, e g , Solow 1963, p. 98) has emphasized an alternative "aggregate"—the consumption rate of return (aggregating over different consumption goods rather than over different capital goods). It can be shown, for instance, that if A and B are two technologies which are both viable at a given rate of interest r^* (i.e , they both make zero profits and all other technologies make losses, at the given rate of interest and the associated prices), and if the economy is in steady state in A and wishes to "switch" to steady state B, the present discounted value, at the rate of interest r^*, of the change in consumption (from that on the path remaining at A) is zero, provided that the economy fully utilizes its resources at every date This result does not in any way depend on the ability to form aggregate capital stocks, or on the possibility of reswitching Nor does the fact that in the steady state A there is more of a particular capital good than in steady state B imply that in the transition some of the capital good must become redundant, as Pasinetti (1969) seems to have suggested. (The conditions under which a full-employment transition is possible are far more subtle than that.)

[12] The other comparative dynamic propositions which have been shown not to be generally valid are: (*a*) economies with higher interest rates have lower values of steady-state consumption, and (*b*) economies with higher interest rates have lower aggregate capital labor ratios. The "perverse" case can occur even without reswitching, and, indeed, it can be shown that it may occur even in a mild modification of Wicksell's model See Stiglitz (1973*a*).

There are other issues in capital and growth theory which are not discussed in this book, and these omissions are perhaps as revealing as the topics which are discussed For instance, as I have noted above, most of the analysis concerns steady-state growth paths. The interesting and difficult questions concerning disequilibrium dynamics, expectation formation, and adjustment processes cannot be properly analyzed in that context [13] Second, the models analyzed all ignore the role of the government and, in particular, of the monetary authorities in determing the growth path of the economy. Finally, the vast amount of work in normative capital theory—for example, the characterization of efficient paths and the analysis of optimal growth paths—is completely ignored. Even if one doubts the validity of the neoclassical models as a description of how competitive economies develop, the models may prove to be of some use in planned economies

It is difficult to leave this subject without asking, How can we explain the Cambridge-Cambridge controversy? This is a question which a sociologist of knowledge might perhaps be in the best position to answer Although "personalities" clearly have something to do with the dispute, the fact that the controversy has received such widespread interest suggests that there is more to it than just that. I suggest that there are four reasons—all well illustrated in this book—why a seemingly dry subject like capital theory is the subject of such controversy

1 Harcourt and the Cambridge (U K) economists keep insisting that adherence to the alternative approaches to capital theory is a matter of ideology As Harcourt puts it "It is my strong impression that if one were to be told whether an economist was fundamentally sympathetic or hostile to basic capitalist institutions, especially private property and the rights to income streams, or whether he were a hawk or a dove in his views on the Vietnam War, one could predict with a considerable degree of accuracy . which side he would be on in the present controversies " Without examining the empirical validity of the statement,[14] one can see that it illustrates the common confusion in the Cambridge (U K.) approach between correlation and causation For Harcourt and the Cambridge (U K) economists, one cannot separate ideology from economic analysis, just as one cannot separate the determinants of the rewards to factors from the analysis of the personal distribution of income

---

[13] For instance, the differences in stability properties of the Harrod and Solow models and the questions of stability of economies with heterogeneous capital goods (see Hahn 1966) are never discussed. The question of whether "savings determines investment or investment determines savings" is also a question which cannot be properly analyzed in a steady-state model with full employment, for in that context, an increase in savings must necessarily be associated with an increase in investment.

[14] As I note below, one of the chief characteristics of the Cambridge (U.K.) "methodology" is a heavy reliance on "casual empiricism" and a reluctance to subject alternative hypotheses to statistical verification

and the patterns of expenditure of that income Many other economists would argue that ideology plays a far less important role in their economic analysis than Harcourt would suggest. It may have some limited influence on the questions we ask, but not on how we go about answering them, for instance, the F-test is not ideological, nor is the logical consistency of the assumptions of neoclassical analysis an ideological issue.

2. There has been a remarkable absence of an attempt at empirical verification of any of the underlying hypotheses at any but the most casual level by advocates of the Cambridge (U K) approach Certainly some of the underlying hypotheses should be testable (e g , the alternative-savings hypotheses) In fact, on many of the issues, the economic data we have may not be adequate to say that one approach is much better than another, yet a proper account of the alternative approaches would have made clear what tests should be performed

3 What makes the matter more difficult is not only the absence of any attempt at empirical verification with any but the crudest of "casual empiricism" but a fundamental misunderstanding of the role of economic theory A model which is appropriate for illuminating one class of problems may not be adequate for illuminating another we know that Newtonian physics is only a first approximation, that gravity does not quite vary with the square of the distance, etc Yet for most of the questions we wish to discuss in elementary physics courses, it is perfectly adequate Similarly, an aggregative growth model may do reasonably well in explaining long-term movements in certain macroeconomic variables

4 There is a well-known propensity of individuals to dislike what they don't or can't understand This book, as well as the writings of the other Cambridge economists, makes perfectly clear that they do not understand neoclassical capital theory A modern treatise on capital theory which makes no reference to Malinvaud, to Debreu, or to Samuelson's consumption loan model is indeed striking To a neoclassical economists (corn, money,) interest rates are just intertemporal prices If p_c is the price today of corn delivered at time $t + 1$, then the corn interest rate is just $[p_c(t + 1)]/[p_c(t)] - 1$ The relationship between prices (interest rates) and marginal products, in a competitive economy, is fundamentally the same in the capital theoretic models as it is in the static models One can object to the competitive assumption, one can say that firms use markup pricing (without explaining how the markups are determined, except vague references to competitive forces), but these are not controversies in capital theory Indeed, most economists would argue that the competitive profit-maximizing model is a more appropriate description for the behavior of the economy in the long run than it is in the short run; that is, there is more competition in the long run than in the short, and in the long run, firms that happen to maximize profits will

do better than firms that don't (the Schumpeterian approach). Thus, the appropriateness of the marginal-productivity theory as a theory of distribution of income among factors is completely unrelated to any of the controversies concerning double switching, savings behavior, or the aggregation of capital

(There is one important difference between the "static" model and the intertemporal model rising from the almost complete absence of futures markets This means that an analysis of expectation formation is crucial in the description of dynamic economies. To their credit, the Cambridge [U K] economists have stressed the role of expectations and stochastic disturbances, but unfortunately the role of expectation formation is obscured by focusing on "Golden Ages" [steady-state analysis] as the Cambridge economists—as well as most others—have done.)

Yet it appears that it is the confused attempt to discredit the marginal-productivity interpretation of the interest rate which imbues the topics of capital theory with their ideological interest to the devotees of Cambridge (U.K) doctrine

Even though Harcourt's book will not lead to a resolution of the issues, or even to a resolution of what are the real issues, by setting out in perhaps as complete and intelligible a manner as is possible one side's view of the debate, Harcourt may have performed a service

References

Cass, D , and Yaari, M. E "Individual Saving, Aggregate Capital Accumulation, and Efficient Growth." In *Essays on the Theory of Economic Growth*, edited by K Shell. Cambridge, Mass. M.I T. Press, 1967.

Diamond, P A. "National Debt in a Neoclassical Growth Model." *A.E R* 55, no. 5, pt. 1 (December 1965) 1126–50.

Hahn, F H "Equilibrium Dynamics with Heterogeneous Capital Goods " *Q.J E* 80, no. 4 (November 1966) 633–46.

Harcourt, G. C "Some Cambridge Controversies in the Theory of Capital " *J. Econ. Literature* 7, no 2 (June 1969) 369–405

Pasinetti, L. L. "Rate of Profit and Income Distribution in Relation to the Rate of Economic Growth." *Rev. Econ. Studies* 29 (October 1962): 267–79.

——— "Switches of Technique and the 'Rate of Return' in Capital Theory " *Econ. J.* 79 (September 1969). 508–31.

Samuelson, P. A. "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money." *J.P.E.* 66, no. 6 (December 1958) 467–82.

———. "A Summing Up " *Q.J.E.* 80, no. 4 (November 1966): 568–83.

Solow, R M. *Capital Theory and the Rate of Return.* Amsterdam · North-Holland, 1963.

Stiglitz, J. E. "Factor Price Equalization in a Dynamic Economy." *J.P.E* 78, no. 3 (May/June 1970): 456–88.

———. "The Badly Behaved Economy with the Well Behaved Production " In *Models of Economic Growth*, edited by J. Mirrlees and N. Stern. Proceeding of the International Economic Association conference, Jerusalem, 1970 London: Macmillan, 1973 *(a)*

———. "Recurrence of Techniques in a Dynamic Economy" In *Models of Economic Growth*, edited by J. Mirrlees and N. Stern. Proceedings of the International Economic Association conference, Jerusalem, 1970. London: Macmillan, 1973. *(b)*

do better than firms that don't (the Schumpeterian approach). Thus, the appropriateness of the marginal-productivity theory as a theory of distribution of income among factors is completely unrelated to any of the controversies concerning double switching, savings behavior, or the aggregation of capital.

(There is one important difference between the "static" model and the intertemporal model rising from the almost complete absence of futures markets. This means that an analysis of expectation formation is crucial in the description of dynamic economies To their credit, the Cambridge [U K] economists have stressed the role of expectations and stochastic disturbances, but unfortunately the role of expectation formation is obscured by focusing on "Golden Ages" [steady-state analysis] as the Cambridge economists—as well as most others—have done.)

Yet it appears that it is the confused attempt to discredit the marginal-productivity interpretation of the interest rate which imbues the topics of capital theory with their ideological interest to the devotees of Cambridge (U K) doctrine

Even though Harcourt's book will not lead to a resolution of the issues, or even to a resolution of what are the real issues, by setting out in perhaps as complete and intelligible a manner as is possible one side's view of the debate, Harcourt may have performed a service

References

Cass, D , and Yaari, M E. "Individual Saving, Aggregate Capital Accumulation, and Efficient Growth." In *Essays on the Theory of Economic Growth,* edited by K Shell. Cambridge, Mass. M.I.T. Press, 1967.

Diamond, P. A. "National Debt in a Neoclassical Growth Model." *A.E R* 55, no. 5, pt. 1 (December 1965) 1126–50.

Hahn, F H. "Equilibrium Dynamics with Heterogeneous Capital Goods" *Q.J E* 80, no 4 (November 1966) 633–46

Harcourt, G. C. "Some Cambridge Controversies in the Theory of Capital" *J. Econ. Literature* 7, no. 2 (June 1969). 369–405

Pasinetti, L. L. "Rate of Profit and Income Distribution in Relation to the Rate of Economic Growth." *Rev. Econ. Studies* 29 (October 1962): 267–79.

———. "Switches of Technique and the 'Rate of Return' in Capital Theory" *Econ. J.* 79 (September 1969) · 508–31.

Samuelson, P. A. "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money." *J.P.E.* 66, no. 6 (December 1958) 467–82.

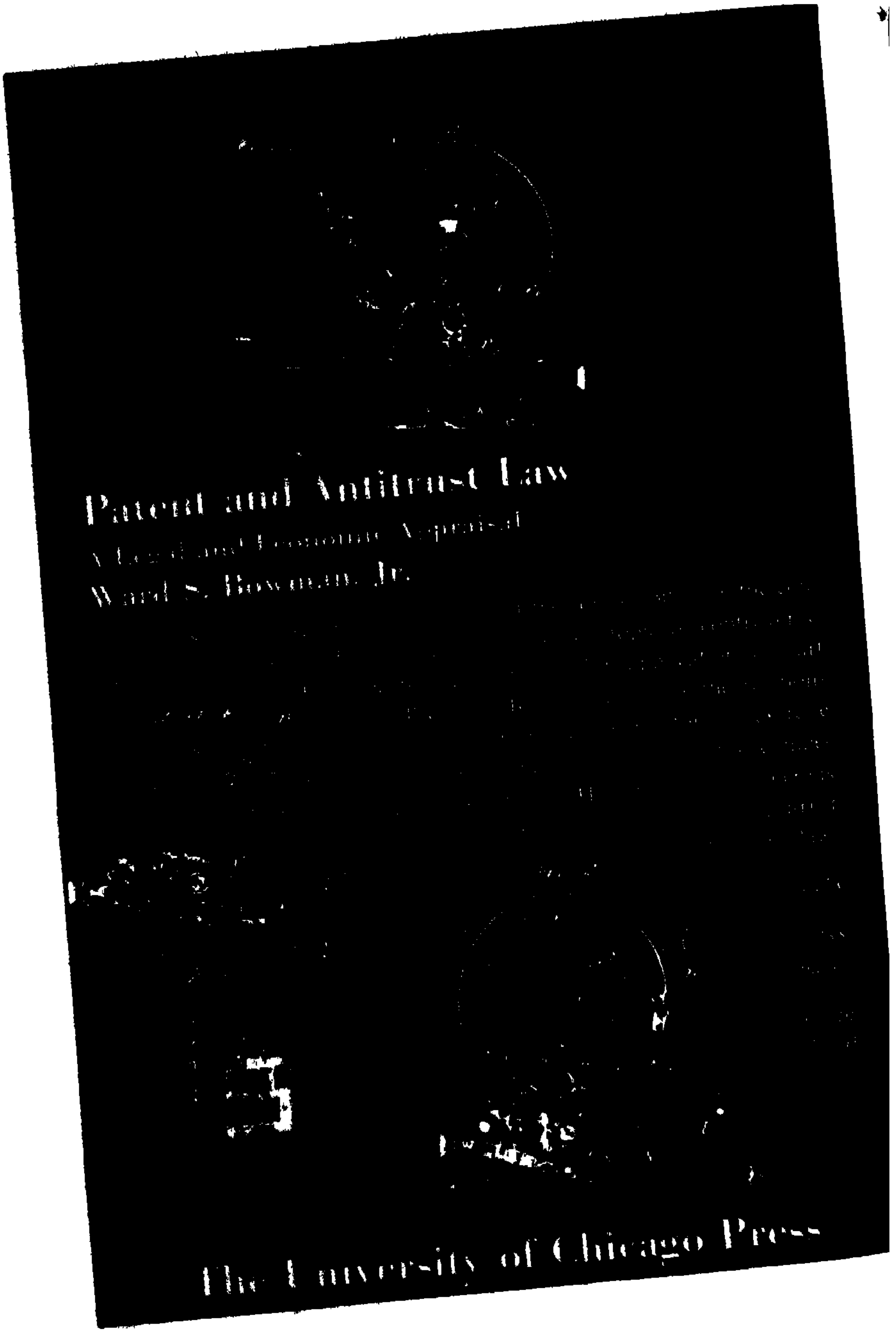
———. "A Summing Up." *Q.J.E.* 80, no. 4 (November 1966): 568–83

Solow, R. M. *Capital Theory and the Rate of Return.* Amsterdam: North-Holland, 1963.

Stiglitz, J. E. "Factor Price Equalization in a Dynamic Economy." *J.P E* 78, no. 3 (May/June 1970): 456–88.

———. "The Badly Behaved Economy with the Well Behaved Production" In *Models of Economic Growth,* edited by J. Mirrlees and N. Stern. Proceedings of the International Economic Association conference, Jerusalem, 1970 London: Macmillan, 1973. *(a)*

———. "Recurrence of Techniques in a Dynamic Economy." In *Models of Economic Growth*, edited by J. Mirrlees and N. Stern. Proceedings of the International Economic Association conference, Jerusalem, 1970 London: Macmillan, 1973. *(b)*



Patent and Antitrust Law

A Legal and Economic Appraisal

Ward S. Bowman, Jr.

The University of Chicago Press

Journal of Political Economy

Volume 82 Number 2, Part II March/April 1974

S1 Prefatory Note

S2 The High Value of Human Time: Population Equilibrium
Theodore W Schultz

S11 A Theory of Marriage Part II
Gary S Becker

S27 Comment The Economics of Nonmonetary Variables
William J Goode

S34 The United States Marriage Market
Alan Freiden

S54 Comment
T Dudley Wallace

S57 Benefits of Women's Education within Marriage
Lee Benham

S72 Comment
Finis Welch

S76 Family Investments in Human Capital Earnings of Women
Jacob Mincer and Solomon Polachek

S109 Comment
Otis Dudley Duncan

S111 Home Investments in Children
Arleen Leibowitz

S132 Comment
Frank P. Stafford

S136 Effects of Child-Care Programs on Women's Work Effort
James J. Heckman

S164 Comment
Sherwin Rosen

S170 Economics of Postwar Fertility in Japan: Differentials and Trends
Masanori Hashimoto

S195 Comment
Gary R. Saxonhouse

S200 Household and Economy: Toward a New Theory of Population and Economic Growth
Marc Nerlove

S219 Comment
Zvi Griliches

S222 Combined References

National Bureau of Economic Research, 1974

OFFICERS Arthur F Burns, *Honorary Chairman*, Walter W Heller, *Chairman*, J Wilson Newman, *Vice Chairman*, John R Meyer, *President*, Thomas D Flynn, *Treasurer*, Douglas H Eldridge, *Vice President-Executive Secretary*, Victor R Fuchs, *Vice President-Research & Co-director, NBER-West*, Edwin Kuh, *Director, Computer Research Center*, Hal B Lary, *Vice President-Research*, Robert E Lipsey, *Vice President-Research*, Sherman J Maisel, *Co-director, NBER-West*, Geoffrey H Moore, *Vice President-Research*, Edward K Smith, *Vice President*

DIRECTORS AT LARGE Atherton Bean, *International Multifoods Corporation*, Joseph A Beirne, *Communications Workers of America*; Arthur F Burns, *Board of Governors of the Federal Reserve System*, Wallace J Campbell, *Foundation for Cooperative Housing*, Erwin D Canham, *Christian Science Monitor*, Emilio G Collado, *Exxon Corporation*, Solomon Fabricant, *New York University*, Eugene P Foley, *Montrose Securities, Inc* , Eli Goldston, *Eastern Gas and Fuel Associates*, David L Grove, *International Business Machines Corporation*, Walter W Heller, *University of Minnesota*, Vivian W Henderson, *Clark College*, John R Meyer, *Harvard University*, J Irwin Miller, *Cummins Engine Company, Inc* , Geoffrey H Moore, *National Bureau of Economic Research*, J Wilson Newman, *Dun & Bradstreet, Inc* , James J O'Leary, *United States Trust Company of New York*, Alice M Rivlin, *Brookings Institution*, Robert V Roosa, *Brown Brothers Harriman & Co* , Boris Shishkin, *Washington, D C* , Arnold M Soloway, *Jamaicaway Tower, Boston, Massachusetts*, Lazare Teper, *International Ladies' Garment Workers' Union*, Donald B Woodward, *Riverside, Connecticut*, Theodore O Yntema, *Oakland University*

DIRECTORS BY UNIVERSITY APPOINTMENT Moses Abramovitz, *Stanford*, Gardner Ackley, *Michigan*, Charles H Berry, *Princeton*, Francis M Boddy, *Minnesota*, Otto Eckstein, *Harvard*, Walter D Fisher, *Northwestern*, R A Gordon, *California*, Robert J. Lampman, *Wisconsin*, Maurice W Lee, *North Carolina*, Almarin Phillips, *Pennsylvania*, Lloyd G. Reynolds, *Yale*, Robert M Solow, *Massachusetts Institute of Technology*, Henri Theil, *Chicago*, William S Vickrey, *Columbia*, Thomas A. Wilson, *Toronto*

DIRECTORS BY APPOINTMENT OF OTHER ORGANIZATIONS Eugene A Birnbaum, *American Management Association*, Thomas D. Flynn, *American Institute of Certified Public Accountants*; Nathaniel Goldfinger, *American Federation of Labor and Congress of Industrial Organizations*; Harold G Halcrow, *American Agricultural Economics*

Association; Walter E. Hoadley, *American Finance Association*; Philip M. Klutznick, *Committee for Economic Development*; Roy E. Moor, *National Association of Business Economists*; Douglass C. North, *Economic History Association*; Willard L. Thorp, *American Economic Association*; W. Allen Wallis, *American Statistical Association*, Robert M. Will, *Canadian Economics Association*

DIRECTORS EMERITI · Percival F. Brundage, Frank W. Fetter; Gottfried Haberler, Albert J. Hettinger, Jr.; George B Roberts; Murray Shields; Joseph H. Willits.

SENIOR RESEARCH STAFF. Gary S. Becker, Charlotte Boschan; Phillip Cagan, Stanley Diller; Solomon Fabricant; Milton Friedman, Victor R. Fuchs, J. Royce Ginn; Raymond W. Goldsmith, Michael Gort; Michael Grossman, F Thomas Juster, John F Kain, John W Kendrick, Irving B. Kravis; Edwin Kuh, William M Lander, Hal B. Lary, Robert E. Lipsey, Benoit B Mandelbrot, John R Meyer; Robert T. Michael, Jacob Mincer, Ilse Mintz, Geoffrey H. Moore, M. Ishaq Nadiri, Nancy Ruggles, Richard Ruggles, Anna J. Schwartz, Robert P Shay, Edward K Smith, George J. Stigler, Victor Zarnowitz

Other Conference

Relation of the National Bureau Directors to Publications Reporting Conference Proceedings

Since the present volume is a record of conference proceedings, it has been exempted from the rules governing submission of manuscripts to, and critical review by, the Board of Directors of the National Bureau. [*Resolution adopted July 6, 1948, as revised November 21, 1949, and April 20, 1968*]

The current interest in population economics at the National Bureau of Economic Research has been encouraged by a four-year grant from the Ford Foundation The National Bureau wishes to acknowledge with thanks the Ford Foundation's support of this research program and the two population conferences The Population Council's joint sponsorship of these conferences is also gratefully acknowledged.

Prefatory Note

All papers, except the first, appearing in this, the second Supplement to the *Journal of Political Economy* devoted to the economics of fertility, were presented at a conference held June 4 and 5, 1973 The searching and critical discussion that followed the invited comment on each of these papers helped the authors appreciably in the revisions that most of them made prior to publication

The National Bureau of Economic Research and the Population Council are the sponsors and provided the funds for both the 1972 conference, the proceedings of which appeared in the March/April 1973 Supplement, "New Economic Approaches to Fertility," and the 1973 conference, from which this Supplement evolved In planning the 1973 program, I had the advice of Gary S Becker, Victor Fuchs, Robert T Michael, Marc Nerlove, T Paul Schultz, and Robert J Willis

Once again I apologize to those who wished to attend the conference but for whom there was no room because of my decision to keep the number of participants small Others who attended and took part in the conference, in addition to the authors and discussants listed in this Supplement, were Mary Jean Bowman, Harley Browning, Tomas Frejka, Victor Fuchs, H Gregg Lewis, Robert T Michael, Warren Sanderson, T Paul Schultz, Joseph J Spengler, Boone A Turchi, and Robert J Willis

The favorable responses to the combined list of all references, as they were presented in the first Supplement, has led Mrs. Virginia K Thurner and me to repeat the combined listing of all references in this Supplement

As was true last time, the editorial quality of this Supplement owes much to the competent work of Mrs Thurner

THEODORE W SCHULTZ

July 25, 1973

The High Value of Human Time: Population Equilibrium

Theodore W. Schultz

University of Chicago

In the continuing economic dialogue between theory and data about fertility, we see once again that fertility is influenced by the differences in the opportunity costs of children associated with the stock of human capital of mothers. The quality of the sacrifices that parents make in acquiring the satisfactions they derive from children influences the quality of the children. But quality is expensive. On these and other important points, the core of the studies in this volume provides additional evidence supporting the first set of studies (T. W. Schultz 1973*c*). Mincer and Polachek evaluate the on-the-job training that women forego on leaving their jobs to have and to care for their children. Heckman examines the effects of child-care programs on the labor-force work efforts of women. Leibowitz, using a unique sample of children with data on them spanning virtually four decades, provides important new evidence on home investment in children. These studies are complemented by Becker's extensions of his theory of marriage and by Freiden's application of the theory to the U.S. marriage market. Benham advances a hypothesis for interpreting the effects of wives' education on the earnings of their husbands.

Turning to the marked decline in fertility in Japan, Hashimoto shows that much of it is explained by the economic variables that are featured in the household approach to fertility. But the sociologists who participated in the conference—Browning, Duncan, and Goode—challenged us repeatedly with the argument that we are not bold enough in extending and applying our economic theory to fertility.

Admittedly, except for the Taiwan study by T. Paul Schultz (1973) we have been shy in extending our theory to fertility behavior in low

---

I am indebted to Gary S. Becker, Marc Nerlove, and T. Paul Schultz for their helpf comments on an early draft of this paper.

income countries. Then, too, with the exception of Nerlove's paper in this volume, we have been timid in developing a framework for analyzing the interactions between the micro behavior of households and the changing state of the economy.[1] Griliches's comment advances the inquiry, and the searching discussion that followed at the conference indicated clearly the importance of this neglected problem Although boldness is called for in solving this problem, it alone will not suffice

As the title of my paper implies, I shall examine issues that are related to Nerlove's analysis. The latter part of his paper opens an important new vista Since his approach to the high price of human time parallels my recent work, as he notes in his paper, and since it has been in the forefront of my thinking, my aim is to supplement this part of his paper I shall first consider two widely different population equilibrium concepts, in support of the second concept I shall extend the explanation of the secular increases in the value of human time I will then take a critical look at the usefulness of the household model in analyzing fertility in countries characterized by high birth and death rates and by substantial increases in national income in a context where the level of the economic value of human time is very low and not increasing much

I

The idea of a population equilibrium is here viewed strictly as an analytical concept to guide economic thinking in deriving testable propositions pertaining to fertility As an analytical device there is no presumption that any population ever has arrived or will arrive at a precise equilibrium In this respect the concept of a population equilibrium is on a par with the many other equilibrium concepts that abound in economic analysis

Two very different concepts of a population equilibrium can now be formulated They may be viewed as types at the two extremes with respect to the state of the economy The first rests on the proposition that the equilibrium is basically a consequence of increases in the price of the services of natural resources relative to capital and labor (wages) The underlying proposition of the second concept is that the equilibrium is determined predominantly by increases in the price of human time relative to that of materials. The per capita income implication of the first is subsistence and that of the second a high standard of living

The first concept, as it was envisaged by the early English economists, has long been a standard part of economics It assumes that the supply of land is fixed and that diminishing returns gradually increase the price of food as a consequence of population growth Gains in productivity from capital are exhausted by this process. This concept can, of course, be extended to encompass the results of the recent macrosystem models that purport to show the limits of the earth in accommodating population growth These models are not restricted solely by the availability of land to produce food, since they also include the physical limits set by the availability of minerals, energy, and space for people The fertility behavior of people in these models is crudely Malthusian, population growth stops (suddenly) as a consequence of the inevitable food, energy, and space crisis Within the Ricardian framework, this concept is a logical conception of a population equilibrium. It is a dismal view of human behavior that has long been an important idea in social thought.

---

[1] For concern about aspects of this problem, see Yoram Ben-Porath (1973), T Paul Schultz (1973), and Marc Nerlove and T. Paul Schultz (1970)

The foundation of the second concept is the high price of human time relative to the price of the services of material factors and goods The concept rests on the proposition that the state of the economy is such that the economic role of the services of natural resources and of intermediate material products is small relative to the role of the services of human agents in production and in consumption, in the sense that the value of the contribution of materials to human satisfactions is small compared with that of human agents In the context of such an economy, the opportunity cost of bearing children is high and the investment in their human capital is large The welfare implications of this concept are unmistakably optimistic because the gains in productivity from the accumulation of human and nonhuman capital are transformed into high standards of living supported by high per capita income Advances in useful knowledge, embodied in human and nonhuman capital, have gradually destroyed the assumption of the fixed supply of the "original properties of the soil " In the process, it is the scarcity of human time and its high value that dominate, and it is the "fixed supply of human time" consisting of 24 hours per day and of a man's lifetime that becomes the critical factor in analyzing the economic behavior of people, including their fertility

Empirically, there is an abundance of evidence which shows that the price of human time accounts for most of the costs in a modern economy The upward tendency of real wages and salaries, including fringe benefits, of earnings foregone by mature students, and of the value of the time of housewives relative to the price of materials is well documented [2] Economic theory implies, and we observe, that material goods are substituted for human time by firms and by households. Received theory, however, is silent on the effects of the high and rising price of human time on pure consumption, although consumption obviously entails time In my thinking, the ultimate economic limit of affluence (economic growth) is not in the scarcity of material goods but in the scarcity of human time for consumption[3]

---

[2] Evidence, for example, on long-term changes in wages and salaries relative to rent paid for the services of farmland in the United States shows that the total real compensation per hour at work of all manufacturing-production workers increased between 1929 and 1970 more than four times as much as did the rent on farm real estate per acre, similarly adjusted, under the assumption that rents tended to parallel the changes in the price of farm real estate per acre (T W Schultz 1972*b*)

Theoretically, the critical postulate assumes that there is a dynamic process that determines the increases in the price of human time relative to the price of the services of the nonhuman factors and that this process tends toward an equilibrium The dynamic part is the economic key to the following four issues (1) the relative increase in investment in human capital augmenting the quality of human beings, (2) the relatively high price of all labor time–intensive goods and other sources of labor–intensive satisfactions, including children, thus leading to the substitution of quality for numbers of children, (3) the relatively cheap material goods that are not labor-intensive, and (4) the scarcity of the time for consumption, setting the ultimate limit to the satisfactions that can be derived from materials provided by economic growth

Although it is obvious that the economic value of human time is high in the affluent countries that have a modern economy, it is not obvious why these economies have developed the demand for and supply of human abilities that have such a high value in terms of earnings and satisfactions that people derive from them I shall elaborate on these developments (see T W Schultz 1972*b*, 1973*a*) and examine some of their basic aspects

My approach to the persistent secular increase in the economic value of human time consists of a highly simplified framework to get at the supply and demand developments that appear to be determining the rise of the price of human time in the context of the modernizing processes The developments explaining the increases in the supply of the quality attributes of human agents are fairly clear, whereas the developments underlying the increases in the demand for these quality attributes are still not clear Recent advances in economic analysis (summarized in T W Schultz 1972*a* and 1973*b*) provide the major parts of the theory for determining the supply of these quality attributes They treat the useful abilities that people acquire as forms of human capital The investment in these abilities is taken to be in response to favorable investment opportunities, and thus the increases in the supply depend on current expenditures (sacrifices) made by individuals, by families, and by public bodies on education, health, job training, the search for information, and geographical migration to take advantage of better jobs or of better consumption opportunities These expenditures (sacrifices) are

---

[3] The approach outlined in this paragraph and a considerable part of the argument that follows appear in my Woody Thompson lecture to the Midwest Economics Association, "Explanation and Interpretations of the Increasing Value of Human Time" (T W Schultz 1973*a*)

[illegible] as moderate-
[illegible] implications that
[illegible] value of human time
[illegible] resources.[7] It implies that the
[illegible] to labor increases over time.[8]
[illegible] the allocative ability of both

[illegible] and mine.
[illegible] appears in previous works
[illegible] 1 and 2 in my *Investment in*
[illegible] 12 treats the "Allocation

[illegible] cost of investment resources
[illegible] long-term real rate of interest
[illegible] capital embodied in human
[illegible] with the going rate of

[illegible] size relative to income

presumably made deliberately with an eye to future satisfactions and earnings The theory of the allocation of time and the household production model are of special importance in analyzing the incentives and responses of people in acquiring education and job training, in enhancing their health, in searching for information, and in altering their fertility, including the substitution of quality for numbers of children. Thus, these supply responses to the increases in the economic incentives associated with modernization are not hard to comprehend The human capital literature abounds with studies dealing with aspects of these supply responses

But these human capital studies have not explained the secular increases in the demand for these quality attributes of human agents The clue to this unresolved puzzle is concealed in two basic factual issues The first of these is that diminishing returns to capital have not occurred generally, despite the vast accumulation of capital in the advanced economies The second is the relatively high rate at which the formation of human capital has occurred Of the two, the first is fundamental, and the resolution of it provides a solution for the second In my thinking, the key to both is in that part of the economic process that increases the stock of useful knowledge.[4] It is the acquisition, adoption, and efficient utilization of this knowledge that have provided *the decisive new sources of investment opportunities* that have maintained the growth process and have kept the returns to capital from diminishing over time Furthermore, these additions to the stock of knowledge have been relatively more favorable in increasing the investment opportunities in the quality attributes of human agents than in the quality components of material agents of production The investment incentives that are revealed by the inequalities in these investment opportunities, as they occur over time, are the mainspring in this process

In an all-inclusive view of these investment opportunities, the knowledge-producing sector must also be included It is not a trivial sector in modern countries, nor is it exogenous Research is an organized activity that requires specific, expensive, scarce resources. Although research is costly, recent studies, most of them devoted to analyzing the rates of return to investment in organized agricultural research, show very high social rates of return

With respect to this investment process, economists could have been spared much aimless wandering had they perceived the implications of the concept of capital as Marshall sketched it for us. His predecessors had formulated the concept of the "state of the productive arts," and they then proceeded to develop the core of economic theory under the assumption that these arts remained constant It was an ingenious simplification, and their theory was in general relevant to a wide array of problems of their day But industrialization undermined this simplifying assumption, and Marshall saw it clearly and cogently. In his treatment of the agents of production, he extended the concept of labor to include work with our hands and our heads It should be noted with care that his concept of capital "consists in great part of knowledge and organization, and of this some part is private property and the other part is not *Knowledge is our most powerful engine of production* Organization aids knowledge The distinction between public and private property in knowledge and organization is of great and growing importance in some respects of more importance than that between public and private property in material things "[5] In not seeing the implications of Marshall's remarkable insights, economists have wandered for years in the wilderness of capital confined to material goods

---

[4] Simon Kuznets in his Nobel Prize lecture, which appeared in the June 1973 *American Economic Review* under the title "Modern Economic Growth· Findings and Reflections" (Kuznets 1973), also attributes a major role to the additions in knowledge in this context He argues that the last two centuries have been periods during which there has occurred "enormous accumulation in the contribution to the stock of useful knowledge by basic and applied research" (p. 251)

Thus, in a nutshell, the persistent increase in the demand for the high-quality services of human agents is a function of the additions to the stock of useful knowledge [6] The complexities of the additions to this knowledge have been much greater in recent, modern economic growth than during early, relatively simple industrialization The rate at which the stock of useful knowledge has increased has also been higher than the rate at which it grew during the early stages of industrialization

This approach has broad integrative power in that it provides a unifying principle for a consistent explanation of the allocation of investment resources encompassing both human and nonhuman capital as modernization proceeds From it we derive important empirical implications that can be tested against data It implies that the value of human time increases relative to the cost of investment resources [7] It implies that the relative share of national income accruing to labor increases over time [8] It implies that there is a special premium for the allocative ability of both males and females in managing firms[9] and households and in allocating their own time, including investments in themselves. It also implies that as the value of the time of mothers increases, fertility declines.[10] These implications are derived from the dynamic process, not from a model of the economy that has arrived at a general equilibrium.

---

[5] Marshall (1930, bk 4, chap 1, pp 138–39), the italics are mine

[6] The argument in support of this summary statement appears in previous works (T W Schultz 1972*b*, 1973*a*) It is anticipated in chaps 1 and 2 in my *Investment in Human Capital The Role of Education and Research* (1971) Chapter 12 treats the "Allocation of Resources to Research "

[7] A simplified approach to this implication is to treat the cost of investment resources as constant under the assumptions that the "normal" long-term real rate of interest remains constant and that, from an increasing amount of capital embodied in human beings, people derive earnings and satisfactions commensurate with the going rate of interest

[8] As the earnings from the increasing stock of human capital rise relative to income acquired from property assets

The concept of a general economic equilibrium in this context is necessary, however, as an analytical guide. It is an assumed economic state toward which this modernization process tends Given this state, there are no inequalities among investment opportunities. The high price of human time is stable in the sense that it is no longer increasing relative to other factor service prices. There is no incentive to make additional investments in human capital or in the knowledge-producing sector, as a consequence of the completion of the modernization process, and advances in knowledge no longer augment the productivity of human time within firms and households, presumably, virtually all of the value added in production is contributed by the input of human time The basic economic constraint that determines the upper limits of modernization (economic growth) in this equilibrium model is the increasing scarcity of human time for consumption The underlying logic can be put very simply modernization increases the consumption stream, consumption requires human time, advances in knowledge, whether they are embodied in material capital or in human capital, are severely limited in the extent to which they can alleviate the scarcity of human time for consumption

For the purpose at hand, the bearing and rearing of children is a very labor-intensive activity, the satisfactions that parents derive from their children is a large part of their "standard of living," and the process of enjoying these satisfactions requires much time, the economic value of which, in this context, is very high

II

Turning to fertility behavior in the low-income countries, the household model as it now stands has not been developed to treat the particular classes of circumstances that constrain the household in these countries These are countries in which illiteracy abounds, human time is cheap, and the income opportunities that women have outside the home are mainly not jobs in the labor market Furthermore, infant mortality is high, life expectancy at birth is low, debilitation during the adult years is substantial for reasons of inadequate nutrition and endemic diseases, and the availability of modern contraceptive techniques, including information about them, is, in general, wanting These classes of circumstances are not as yet at home in the household model

---

[9] As a consequence of the dynamics of the economy, the premium for allocative ability, distinguished from the ability of people to do useful work, is substantial (see Welch 1970, Fane 1972, Huffman 1972, Khaldi 1972)

[10] Nerlove's subtle and incisive treatment of the interactions between the high value of time and fertility lays the groundwork for new thinking with respect to this important implication

The difficulty here is not that economic theory is pointless in explaining fertility behavior in the low-income countries On the contrary, in principle basic economic thinking is fully as applicable to the poor as it is to the rich countries As a case in point, I have long argued (T W Schultz 1964) that the theory of the firm is analytically as powerful in the allocation of resources of poor, small, illiterate farmers in the less-developed countries as it is in determining the allocative efficiency of farmers, say, in Iowa. The usefulness of this theory is now widely recognized because of many recent successful applications The same argument holds for a fully developed theory of the household As yet, however, the part of this theory that has been applied to the United States, Israel, and Japan is a special and narrow part of what I envisage as a general theory of the household Once the additional parts of this theory have been formulated, its usefulness in analyzing household activities in low-income countries will not be in doubt, assuming that it is also extended to treat the effects of economy-wide dynamic development

Households in low-income countries perform, in fact, a substantially larger economic role than they do in high-income countries The value of home production is not only large relative to the total family income, it is also produced predominantly by family labor and only in small part by purchased inputs, because in low-income countries the purchased material goods that households can acquire are very high in price relative to the economic value of the time of members of the household With regard to the costs of children, children are labor-intensive during their infancy regardless of the country's level of income In low-income countries, however, the mother's time is cheap in a context where health services, nutrition, and education for the children are dear The satisfactions and producer services that parents derive from their children are in large measure from an assured number of children to provide help for household work and for family endeavors consisting mostly of farm work and to provide food and shelter for the parents during their old age and only in small measure from human capital that enhances the acquired quality of children. These particular economic constraints on the household are not treated in the part of the theory that has been applied in these fertility studies restricted to rich countries

As an overview, the interactions between the changes in the economy and the economic role of the household including fertility are probably more important in most of these developing countries than in the advanced countries. With respect to which of the two concepts of a population equilibrium is the more relevant in guiding our economic thinking in explaining the fertility behavior in the low-income countries, my view is that it is the second concept My reason for opting for this concept is that the now-popular doomsday literature is not a valid characterization of the direction toward which the economy of such countries is moving. The crude Malthusian view of fertility does not apply, these countries are not headed toward a population equilibrium that is being imposed on them by diminishing returns from natural resources. Per capita income is in general not falling in these countries Measured in terms of living conditions, there are appreciable gains, as is evident from improvements in health and from longer life expectancy Moreover, birth rates are falling and substantially so in an increasing number of low-income countries. Although they are far from arriving at the second type of population equilibrium, they are moving toward it, and it is therefore analytically the relevant concept

Earlier (1973*c*), I expressed my concern that our estimates of the price and income effects on fertility may be subject to considerable error for reasons of changes in the economy over time, changes that are not taken into account in these studies The changes that are taking place in the low-income countries appear to me to be even more important than those in the developed countries, and if this is true, it will be imperative to extend our economic approaches in ways that will make these changes an integral part of the theory and its application. The most pervasive change is in the improvement of health Estimates of life expectancy from time of birth are rising at a high rate relative to the further rise in the rich countries The decline in infant mortality must have important fertility implications Along with somewhat better nutrition, the marked decline in debilitating diseases during adulthood must have some price and income effects on household activities, including fertility The fact that in some parts of poor countries, in the Punjab of India, for example, many girls are now enrolled in schools is an important change in this context Then, too, there is some progress in increasing the supply of information about modern contraceptive techniques and in subsidizing the supply of contraceptives While all of these changes are occurring, the quality of the labor force is slowly rising as is the value of human time

One measure of the fruitfulness of the 1972 and 1973 conferences on the economics of fertility is in the agenda of unfinished research that the papers, comments, and discussions produced There is a couplet by Robert Frost bearing on research Let me paraphrase it

> We sit around the circle and suppose
> The Secret sits in the middle and knows

A Theory of Marriage: Part II

Gary S. Becker

University of Chicago and National Bureau of Economic Research

1. Introduction

In "A Theory of Marriage: Part I," published in the July/August 1973 *Journal of Political Economy*, I presented an analysis of the "marriage market." This paper extends the analysis in that paper to include caring between mates, polygamous marital arrangements, genetic selection related to assortive mating, and separation, divorce, and remarriage. Its purpose is both to enrich the discussion in Part I and to show the power of this approach in handling different kinds of marital behavior.

In Part I (1973) I offered a simplified model of marriage that relies on two basic assumptions: (1) each person tries to find a mate who maximizes his or her well-being, with well-being measured by the consumption of household-produced commodities, and (2) the "marriage market" is assumed to be in equilibrium, in the sense that no person could change mates and become better off. I argued that the gain from marriage compared to remaining single for any two persons is positively related to their incomes, the relative difference in their wage rates, and the level of nonmarket-productivity-augmenting variables, such as education or beauty.

The optimal association between mates with respect to different traits, such as ability, education, race, income, and height, was analyzed. I showed that positive associations, matings of likes, are usually optimal, although with respect to some traits mating of unlikes is optimal, for example, with wage rates. The division of the total "output" produced

---

I am indebted for helpful comments to Isaac Ehrlich, Robert T. Michael, Richard Posner, T. W. Schultz, George J. Stigler, and members of the Workshop in Applications of Economics at the University of Chicago. This research has been supported by a grant from the Ford Foundation to the National Bureau of Economic Research for the study of the economics of population. This paper is not an official NBER publication, since it has not been reviewed by the NBER Board of Directors.

by mates is not assumed to be given a priori, but is determined by the nature of the marriage-market equilibrium.

The simplified analyses in Part I have now been extended in several directions The effect of "love" and caring between mates on the nature of equilibrium in the marriage market is considered Polygamy is discussed, and especially the relation between its incidence and the degree of inequality among men and the inequality in the number of men and women The implications of different sorting patterns for inequality in family resources and genetic natural selection are explored The assumption of complete information about all potential mates is dropped and I consider the search for information through dating, coeducational schools, "trial" marriages, and other ways This search is put in a life-cycle context that includes marriage, having children, sometimes separation and divorce, remarriage, and so forth

2. Love, Caring, and Marriage

In Part I, I ignored "love," that cause of marriage glorified in the American culture At an abstract level, love and other emotional attachments, such as sexual activity or frequent close contact with a particular person, can be considered particular nonmarketable household commodities, and nothing much need be added to the analysis, in the earlier paper, of the demand for commodities That is, if an important set of commodities produced by households results from "love," the sorting of mates that maximizes total commodity output over all marriages is partly determined by the sorting that maximizes the output of these commodities The whole discussion in Part I (1973) would continue to be relevant

There is a considerable literature on the effect of different variables such as personality, physical appearance, education, or intelligence, on the likelihood of different persons loving each other. Since I do not have anything to add to the explanation of whether or why one person would love another, my discussion concentrates on some effects of love on marriage. In particular, since loving someone usually involves caring about what happens to him or her,[1] I concentrate on working out several implications, for marriage, of "caring "

An inclusive measure of "what happens" is given by the level of commodity consumption, and the natural way for an economist to measure "caring" is through the utility function [2] That is, if M cares about F, M's utility would depend on the commodity consumption of F as well as on his own, graphically, M's indifference curves in figure 1 are negatively

---

[1] The *Random House Dictionary of the English Language* includes in its definitions of love, "affectionate concern for the well-being of others," and "the profoundly tender or passionate affection for a person of the opposite sex."

[2] This formulation is taken from my paper, "A Theory of Social Interactions" (1969)

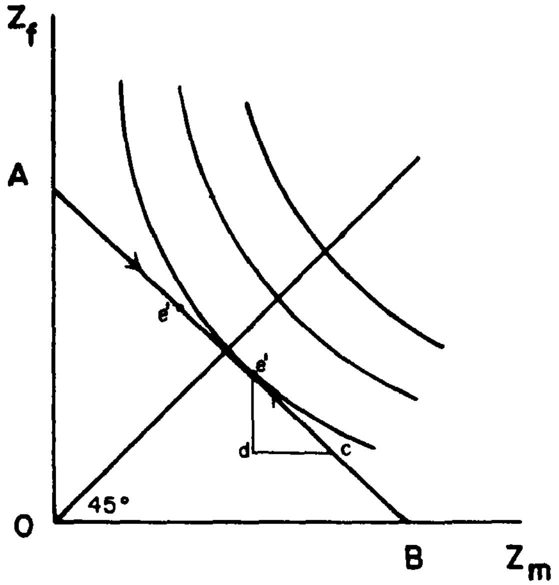


FIG 1

inclined with respect to Z_m and Z_f, the commodities consumed by M and F respectively [3] If M cared as much about F as about himself (I call this "full" caring), the slopes of all the indifference curves would equal unity (in absolute value) along the 45° line,[4] if he cared more about himself, the slopes would exceed unity, and conversely if he cared more about F

Point *c* in figure 1 represents the allocation of commodities to M and F that is determined by equilibrium in the marriage market Only if M were married to F could he transfer commodities to F, since household commodities are transferable within but not between households If the terms of transfer are measured by the line *AB*, he moves along *AB* to point *e* he transfers *cd* and F receives *de* Presumably commodities can be transferred within a household without loss, so that *AB* would have a slope of unity. Then the equilibrium position after the transfer would be on the

---

[3] Since there is only a single aggregate commodity, saying that M's utility depends on F's consumption is equivalent to saying that M's utility depends on F's utility (assuming that F does not care about M) If many commodities $Z_1, \quad , Z_q$, were consumed, M's utility would depend on F's utility if $U^m = U^m[Z_{1m}, \quad , Z_{qm}, g(Z_{1f}, \quad , Z_{qf})]$ where g describes the indifference surface of F Hence $(\partial U^m/\partial Z_{if})/(\partial U^m/\partial Z_{jf}) = (\partial g/\partial Z_{if})/(\partial g/\partial Z_{jf})$; this ratio is F's marginal rate of substitution between Z_i and Z_j

[4] "Full" caring might also imply that the indifference curves were straight lines with a slope of unity, that Z_f was a perfect substitute for Z_m

45° line with full caring, and to the right of this line if M preferred his own consumption to F's.

Most people no doubt find the concept of a market allocation of commodities to beloved mates strange and unrealistic. And, as we have seen, caring can strikingly modify the market allocation between married persons. For example, the final allocation (point e) after the transfer from M to F has more equal shares than does the market allocation (point c).[5] Moreover, if F also cared about M, she would modify the market allocation by transferring resources to M from anywhere in the interval Ae' until she reached a point e',[6] generally to the left of e. The market completely determines the division of output only in the interval $e'e$; positions in Be are modified to e, and those in Ae' are modified to e'. Furthermore, if each fully cares for the other, points e and e' are identical and on the 45° line. Then the total amount produced by M and F would be shared equally, regardless of the market-determined division. This concept of caring between married persons, therefore, does imply sharing—equal sharing when the caring is full and mutual—and is thus consistent with the popular belief that persons in love "share."

Sharing implies that changes in the sex ratio or other variables considered in section 4 of Part I (1973) would not modify the actual distribution of output between married M and F (unless the market-mandated distribution were in the interval ee'). This is another empirical implication of caring that can be used to determine its importance.

I indicated in the earlier paper that total income would be less than total output in a marriage if resources were spent "policing" the market-mandated division of output, whereas total income would exceed total output if some output were a "family" commodity, that is, were consumed by both mates. Caring raises total income relative to total output both by reducing policing costs and by increasing the importance of family commodities.

Consider first the effect of caring on policing costs. "Policing" reduces the probability that a mate shirks duties or appropriates more output than is mandated by the equilibrium in the marriage market.[7] Caring reduces the need for policing. M's incentive to "steal" from his mate F is weaker if M cares about F because a reduction in F's consumption also lowers M's utility. Indeed, caring often completely eliminates the incentive to "steal" and thus the need to police. Thus, at point e in the figure, M has no incentive to "steal" from F because a movement to the right along AB would lower M's utility.[8] Therefore, if M cares about F sufficiently to transfer commodities to her, F would not need to "police" M's consumption[9] Consequently, marriages with caring would have fewer resources spent on "policing" (via allowances or separate checking accounts?) than other marriages would

---

[5] Provided it were in the interval Ae, M would not modify the market allocation.

[6] We assume that AB also gives the terms of transfer for F, and that e' is the point of tangency between AB and her indifference curves.

[7] Policing is necessary in any partnership or corporation, or, more generally, in any cooperative activity (see Becker 1971, pp. 122–23; Alchian and Demsetz 1972).

[8] A fortiori, a movement along any steeper line—the difference between AB and this line measuring the resources used up in "stealing"—would also lower M's utility.

M's income at e exceeds his own consumption because of the utility he gets from F's consumption Indeed, his income is the sum of his and F's consumption, and equals OB (or OA), the output produced by M and F Similarly, F's income exceeds her own consumption if she benefits from M's consumption[10] Caring makes family income greater than family output because some output is jointly consumed At point e, all of F's and part of M's consumption would be jointly consumed Since both e and e' are on the 45° line with mutual and full caring, the combined incomes of M and F would then be double their combined output all of M's and all of F's consumption would be jointly consumed

Love and caring between two persons increase their chances of being married to each other in the optimal sorting That love and caring cannot reduce these chances can be seen by assuming that they would be married to each other in the optimal sorting even if they did not love and care for each other. Then they must also be married to each other in the optimal sorting if they do love and care for each other because love raises commodity output and caring raises their total income by making part of their output a "family" commodity Hence, their incomes when there is love and caring exceed their incomes when there is not Consider the following matrix of outputs

$$\begin{array}{c} \\ M_1 \\ \\ M_2 \\ \\ \end{array} \begin{array}{cc} F_1 & F_2 \\ \end{array}$$

$$\begin{array}{c|cc|} & F_1 & F_2 \\ M_1 & 8 & 4 \\ & (3, 5) & \\ M_2 & 9 & 7 \\ & & (5, 2) \end{array} \qquad (1)$$

With no caring, this is also the matrix of total incomes,[11] and M_1F_1 and M_2F_2 would be the optimal sorting if incomes were sufficiently divisible to obtain, say, the division given in parenthesis With mutual and full caring between M_1 and F_1, m'_{11}, the income of M_1, would equal $8 > 3$, and f'_{11}, the income of F_1, would equal $8 > 5$;[12] clearly, M_1 would still be married to F_1 in the optimal sorting

---

[9] With mutual and full caring, neither mate would have to "police." On the other hand, if each cared more about the other than about himself (or herself), at least one of them, say M, would want to transfer resources that would not be accepted Then F would "police" to prevent undesired *transfers from* M This illustrates a rather general principle, namely, that when the degree of caring becomes sufficiently great, behavior becomes similar to that when there is no caring

[10] F's income equals the sum of her consumption and a fraction of M's consumption that is determined by the slope of F's indifference curve at point e See the formulation in section 1 of the Mathematical Appendix

[11] We abstract from other kinds of "family" commodities because they can be analyzed in exactly the same way that caring is.

That love and caring can bring a couple into the optimal sorting is shown by the following matrix of outputs.

$$\begin{array}{c} \\ M_1 \\ \\ M_2 \\ \\ M_3 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} F_1 & F_2 & F_3 \end{array} \\ \left[\begin{array}{ccc} 10 & 6 & 5 \\ (4, 6) & & \\ 9 & 10 & 4 \\ & (6, 4) & \\ 2 & 3 & 10 \\ & & (5, 5) \end{array}\right] \end{array} . \qquad (2)$$

Without love and caring the optimal sorting is M_1F_1, M_2F_2, and M_3F_3, with a set of optimal incomes given in parenthesis If, however, M_1 and F_2 were in love and had mutual and full caring, the optimal sorting would become M_1F_2, M_2F_1, and M_3F_3 because the incomes resulting from this sorting, $m_{12} = f_{21} = k > 6$,[13] and, say, $m_{21} = f_{21} = 4\frac{1}{2}$, and $m_{33} = f_{33} = 5$, can block the sorting along the diagonal.

Does caring per se—that is, as distinguished from love—encourage marriage for example, couldn't M_1 marry F_1 even though he receives utility from F_2's consumption, and even if he wants to transfer resources to F_2? One incentive to combine marriage and caring is that resources are more cheaply transferred within households by assumption, commodities cannot be transferred between households, and goods and time presumably also are more readily transferred within households Moreover, caring partly results from living together,[14] and some couples marry partly because they anticipate the effect of living together on their caring

Since, therefore, caring does encourage (and is encouraged by) marriage, there is a justification for the economist's usual assumption that even a multiperson household has a single well-ordered preference function For, if one member of a household—the "head"—cares enough about all other members to transfer resources to them, this household would act *as if* it maximized the "head's" preference function, even if the preferences of other members are quite different [15]

Output is generally less divisible between mates in marriages with caring than in other marriages[16] because caring makes some output a family commodity, which cannot be divided between mates. One implication of this is that marriages with caring are less likely to be part of the optimal sorting than marriages without caring that have the same total *income* (and thus have a greater total output).[17]

---

[12] The output of love raises these incomes even further

[13] The difference between k and 6 measures the output of love produced by M_1 and F_2.

[14] So does negative caring or "hatred" A significant fraction of all murders and assaults involve members of the same household (see Ehrlich 1970).

[15] For a proof, see section 1 of the Appendix, further discussions can be found in Becker (1969)

[16] See the proof in section 2 of the Appendix.

Another implication is that the optimal sorting of different traits can be significantly affected by caring, even if the degree of caring and the value of a trait are unrelated. Part I (1973) shows that when the division of output is so restricted that each mate receives a given fraction of the output of his or her marriage, beneficial traits are always strongly positively correlated in the optimal sorting. A negative correlation, on the other hand, is sometimes optimal when output is fully divisible. Caring could convert what would be an optimal negative correlation into an optimal positive one because of the restrictions it imposes on the division of output.

For example, assume that a group of men and women differ only in wage rates, and that *each* potential marriage has mutual and full caring, so that the degree of caring is in this case uncorrelated with the level of wage rates, then the optimal correlation between wage rates would be positive, although I showed in Part I (1973) that it is negative when there is no caring.[18] The (small amount of) evidence in that paper indicating that wage rates are negatively correlated suggests, therefore, that caring does not completely determine the choice of marriage mates.

3. Polygamy

Although monagamous unions predominate in the world today, some societies still practice polygamy, and it was common at one time. What determines the incidence of polygamous unions in societies that permit them, and why have they declined in importance over time?

I argued in Part I (1973) that polyandrists—women with several husbands—have been much less common than polygynists—men with several wives—because the father's identity is doubtful under polyandry. Todas of India did practice polyandry, but their ratio of men to women

---

[17] See the example discussed in section 2 of the Appendix.

[18] As an example, let the matrix of outputs from different combinations of wage rates be

$$\begin{array}{ccc} & F_{w_1} & F_{w_2} \\ M_{w_1} & \left[\begin{array}{c} 5 \\ (5, 5) \end{array}\right. & \left.\begin{array}{c} 10 \\ (10, 10) \end{array}\right] \\ M_{w_2} & \left[\begin{array}{c} 12 \\ (12, 12) \end{array}\right. & \left.\begin{array}{c} 15 \\ (15, 15) \end{array}\right] \end{array}$$

If outputs were fully divisible, the optimal sorting would be $M_{w_1}F_{w_2}$ and $M_{w_2}F_{w_1}$, since that maximizes the combined output over all marriages. With mutual and full caring in all marriages, the income of each mate equals the output in his or her marriage; these incomes are given in parentheses. Clearly, the optimal sorting would now be $M_{w_2}F_{w_2}$ and $M_{w_1}F_{w_1}$.

was much above one, largely due to female infanticide.[19] They mitigated the effects of uncertainty about the father by usually having brothers (or other close relatives) marry the same woman.

I showed in Part I (1973) that if all men and all women were identical, if the number of men equaled the number of women, and if there were diminishing returns from adding an additional spouse to a household, then a monogamous sorting would be optimal, and therefore would maximize the total output of commodities over all marriages.[20] If the plausible assumption of diminishing returns is maintained, inequality in various traits among men or in the number of men and women would be needed to explain polygyny.

An excess of women over men has often encouraged the spread of polygyny, with the most obvious examples resulting from wartime deaths of men. Thus, almost all the male population in Paraguay were killed during a war with Argentina, Brazil, and Uruguay in the nineteenth century,[21] and apparently polygyny spread afterward.

Yet, polygyny has occurred even without an excess of women; indeed, the Mormons practiced polygyny on a sizable scale with a slight excess of men.[22] Then inequality among men is crucial.

If the "productivity" of men differs, a polygynous sorting could be optimal, even with constant returns to scale and an equal number of men and women. Total output over all marriages could be greater if a second wife to an able man added more to output than she would add as a first wife to a less able one. Diminishing marginal products of men or women within each household do not rule out that a woman could have a higher marginal product as a second wife in a more productive household than as the sole wife in a less productive household.

Consider, for example, two identical women who would produce 5 units of output if single, and two different men who would each produce 8 and 15 units, respectively, if single. Let the married outputs be 14 and 27 when each man has one wife, and 18 and 35 when each has two.[23] Clearly, total output is greater if the abler man takes two wives and the other remains single than if they both take one wife: $35 + 8 = 43 > 14 + 27 = 41$ If the abler man received, say, 21 units and each wife received, say, 7 units, no one would have any incentive to change mates

---

[19] See Rivers (1906). Whether the infanticide caused polyandry, or the reverse, is not clear.

[20] An optimal sorting has the property that persons not married to each other could not, by marrying, make some better off without making others worse off. I show in Part I (1973) that an optimal sorting maximizes total output of commodities.

[21] After the war, males were only 13 percent of the total population of Paraguay (see *Encyclopaedia Britannica*, 1973 ed., s.v. "Paraguay"). I owe this reference to T. W. Schultz.

[22] See Young (1954, p. 124). The *effective* number of women can exceed the number of men, even with an equal number at each age, if women marry earlier than men and if widowed women remarry. The number of women married at any time would exceed the number of men married because women would be married longer (to different men—they would be sequentially polyandrous!). This apparently was important in Sub-Saharan Africa, where polygyny was common (see Dorjahn 1959).

[23] These numbers imply diminishing marginal products, since $18 - 14 = 4 < 6$, and $35 - 27 = 8 < 12$.

Our analysis implies generally that polygyny would be more frequent among more productive men—such as those with large farms, high positions, and great strength—an implication strongly supported by the evidence on polygyny For example, only about 10–20 percent of the Mormons had more than one wife,[24] and they were the more successful and prominent ones Although 40 percent of the married men in a sample of the Xavante Indians of Brazil were polygynous, "it was the chief and the heads of clans who enjoyed the highest degree of polygyny" (Salzano, Neel, and Maybury-Lewis 1967, p 473) About 35 percent of the married men in Sub-Saharan Africa were polygynous (Dorjahn 1959, pp 98–105), and they were generally the wealthier men Fewer than 10 percent of the married men in Arab countries were polygynous, and they were the more successful, especially in agriculture (Goode 1963, pp 101–4)

I do not have a satisfactory explanation of why polygyny has declined over time in those parts of the world where it was once more common [25] The declines in income inequality and the importance of agriculture presumably have been partly responsible Perhaps the sex ratio has become less favorable, but that seems unlikely, wartime destruction aside Perhaps monogamous societies have superior genetic and even cultural natural selection (see the next section) But since more successful men are more likely to be polygynous, they are more likely to have relatively many children [26] If the factors responsible for success are "inherited," selection over time toward the "abler" might be stronger in polygynous than in monogamous societies I have even heard the argument that Mormons are unusually successful in the United States because of their polygynous past! However, if the wives of polygynous males were not as able, on the average, as the wives of equally able monogamous males, selection could be less favorable in polygynous societies

The decline in polygyny is usually "explained" by religious and legislative strictures against polygyny that are supposedly motivated by a desire to prevent the exploitation of women But the laws that prevent men from taking more than one wife no more benefit women than the laws in South Africa that restrict the ratio of black to white workers (see Wilson 1972, p. 8) benefit blacks Surely, laws against polygyny reduce the "demand" for women, and thereby reduce their share of total household output and increase the share of men [27]

---

[24] Young (1954, p 441) says that "in some communities it ran as high as 20–25 percent of the male heads of families," but Arrington (1958, p 238) says about 10 percent of all Mormon families were polygynous

[25] Polygyny was more common in Islamic and African societies than in Western and Asian ones, although in China and Japan concubines had some of the rights and obligations of wives (see Goode 1963, chap 5)

[26] Salzano, Neel, and Maybury-Lewis (1967, p 486) found evidence among the Xavante Indians of "similar means but significantly greater variance for number of surviving offspring for males whose reproduction is completed than for similar females " This indicates that polygynous males (the more successful ones) have more children than other males.

4. Assortive Mating, Inequality, and Natural Selection

I pointed out in Part I (1973) that positive assortive mating of different traits reduces the variation in these traits between children in the same family (and this is one benefit of such mating) Positive assortive mating also, however, increases the inequality in traits, and thus in commodity income, between families Note that the effects on inequality in commodity and money incomes may be very different, indeed, if wage rates, unlike most other traits, are negatively sorted (as argued in Part I), assortive mating would reduce the inequality in money earnings and increase that in commodity income

Positive sorting of inherited traits, like intelligence, race, or height, also increases the inequality in these traits among children in different families, and increases the correlation between the traits of parents and children (see proofs in Cavalli-Sforza and Bodmer [1971, chap 9]) Moreover, positive sorting, even of noninherited traits such as education, often has the same effect because, for example, educated parents are effective producers of "education-readiness" in their children (see Leibowitz [1972] and the papers by her and Benham in this volume) The result is an increase in the correlation between the commodity incomes of parents and children, and thereby an increase in the inequality in commodity income among families spanning several generations. That is, positive assortive mating has primary responsibility for noncompeting groups and the general importance of the family in determining economic and social position that is so relevant for discussions of investment in human capital and occupational position

Since positive assortive mating increases aggregate commodity income over all families, the level of and inequality in commodity income are affected in different ways Probably outlawing polygyny has reduced the inequality in commodity income among men at the price of reducing aggregate commodity income. Perhaps other restrictions on mating patterns that reduce inequality would be tolerated, but that does not seem likely at present

---

[27] An alternative interpretation of the religious and legislative strictures against polygyny is that they are an early and major example of discrimination *against* women, of a similar mold to the restrictions on their employment in certain occupations, such as the priesthood, or on their ownership of property This hypothesis has been well stated by (of all people!) George Bernard Shaw "Polygamy when tried under modern democratic conditions as by the Mormons, is wrecked by the revolt of the mass of inferior men who are condemned to celibacy by it, for the maternal instinct leads a woman to prefer a tenth share in a first rate man to the exclusive possession of a third rate " See his "Maxims for Revolutionists" appended to *Man and Superman* (Shaw 1930, p. 220) Shaw was preoccupied with celibacy, he has three other maxims on celibacy, one being "any marriage system which condemns a majority of the population to celibacy will be violently wrecked on the pretext that it outrages morality" (1930, p 220)

Since positive assortive mating increases the between-family variance, it increases the potential for genetic natural selection, by a well-known theorem in population genetics.[28] The actual amount of selection depends also on the inheritability of traits, and the relation between the levels of the traits of mates and the number of their surviving children (called "fitness" by geneticists) For example, given the degree of inheritability of intelligence, and a positive (or negative) relation between number of children and average intelligence of parents, the rate of increase (or decrease) per generation in the average intelligence of a population would be directly related to the degree of positive assortive mating by intelligence

Moreover, the degree of assortive mating is not independent of inheritability or of the relation between number of children and parental traits For example, the "cost" of higher-"quality" children may be lower to more-intelligent parents, and this affects the number (as well as quality) of children desired [29] In a subsequent paper I expect to treat more systematically the interaction between the degree of assortive mating and other determinants of the direction and rate of genetic selection

5. Life-Cycle Marital Patterns

Life-cycle dimensions of marital decisions—for instance, when to marry, how long to stay married, when to remarry if divorced or widowed, or how long to stay remarried—have received little attention in my earlier paper or thus far in this one These are intriguing but difficult questions, and only the broad strokes of an analysis can be sketched at this time A separate paper in the not-too-distant future will develop a more detailed empirical as well as theoretical analysis

A convenient, if artificial, way to categorize the decision to marry is to say that a person first decides when to enter the marriage market and then searches for an appropriate mate [30] The age of entry would be earlier the larger the number of children desired, the higher the expected lifetime income, and the lower the level of education.[31]

---

[28] This theorem was proved by Fisher (1958, pp 37–38) and called "the fundamental theorem of natural selection " For a more recent and extensive discussion, see Cavalli-Sforza and Bodmer (1971, sec 6 7)

[29] For a discussion of the interaction between the quantity and quality of children, see Becker and Lewis (1973)

[30] This categorization is made in an important paper by Coale and McNeil, "The Distribution by Age of the Frequency of First Marriage in a Female Cohort" (1973) They show that the frequency distribution of the age at first marriage can be closely fitted in a variety of environments by the convolution of a normal distribution and two or three exponential distributions. The normal distribution is said to represent the distribution of age at entry into the marriage market, and the exponential distributions, the time it takes to find a mate.

Once in the marriage market, a person searches for a mate along the lines specified in the now rather extensive search literature.[32] That is, he searches until the value to him of any expected improvement in the mate he can find is no greater than the cost of his time and other inputs into additional search Some determinants of benefits and costs are of special interest in the context of the marriage market.

Search will be longer the greater the benefits expected from additional search. Since benefits will be greater the longer the expected duration of marriage, people will search more carefully and marry later when they expect to be married longer, for example, when divorce is more difficult or adult death rates are lower. Search may take the form of trial living together, consensual unions, or simply prolonged dating Consequently, when divorce becomes easier, the fraction of persons legally married may actually *increase* because of the effect on the age at marriage Indeed, in Latin America, where divorce is usually impossible, a relatively small fraction of the adult population is legally married because consensual unions are so important (see Kogut 1972), and, in the United States, a smaller fraction of women have been married in those states having more-difficult divorce laws (see Freiden [1972] and his paper in this volume) [33]

Search would also be longer the more variable potential mates were because then the expected gain from additional "sampling" would be greater Hence, other determinants being the same, marriage should generally be later in dynamic, mobile, and diversified societies than in static, homogeneous ones

People marry relatively early when they are lucky in their search They also marry early, however, when they are unduly pessimistic about their prospects of attracting someone better (or unduly optimistic about persons they have already met) Therefore, early marriages contain both lucky and pessimistic persons, while later marriages contain unlucky and optimistic ones

The cost of search differs greatly for different traits the education, income, intelligence, family background, perhaps even the health of persons can be ascertained relatively easily, but their ambition, resiliency under pressure, or potential for growth are ascertained with much greater difficulty [34] The optimal allocation of search expenditures implies that marital decisions would be based on fuller information about more-easily searched traits than about more-difficult-to-search traits. Presumably, therefore, an analysis of sorting that assumes perfect information (as in Part I [1973]) would predict the sorting by more-easily searched traits, such as education, better than the sorting by more-difficult-to-search traits, such as resiliency [35]

---

[31] For a theoretical and empirical study of these and other variables, see Keeley (1973)

[32] The pioneering paper is by Stigler (1961). For more recent developments, see McCall (1970) and Mortensen (1970).

[33] These results are net of differences in income, relative wages, and the sex ratio

[34] In the terminology of Nelson (1970), education, income, and intelligence are "search" traits, whereas resiliency and growth potential are "experience" traits.

Married persons also must make decisions about marriage should they separate or divorce, and if they do, or if widowed, when, if ever, should they remarry? The incentive to separate is smaller the more important are investments that are "specific" to a particular marriage [36] The most obvious and dominant example of marriage-specific investment is children, although knowledge of the habits and attitudes of one's mate is also significant Since specific investments would grow, at least for quite a while, with the duration of marriage, the incentive to separate would tend to decline with duration

The incentive to separate is greater, on the other hand, the more convinced a person becomes that the marriage was a "mistake" This conviction could result from additional information about one's mate or other potential mates (Some "search" goes on, perhaps subconsciously, even while one is married!) If the "mistake" is considered large enough to outweigh the loss in marriage-specific capital, separation and perhaps divorce will follow

The analysis in Part I (1973) predicts sorting patterns in a world with perfect information Presumably, couples who deviate from these patterns because they were unlucky in their search are more likely than others to decide that they made a "mistake" and to separate as additional information is accumulated during marriage If they remarry, they should deviate less from these patterns than in their first marriage For example, couples with relatively large differences in education, intelligence, race, or religion, because they were unlucky searchers, should be more likely to separate,[37] and should have smaller differences when they remarry In the subsequent paper referred to earlier, I plan to develop more systematically the implications of this analysis concerning separation, divorce, and remarriage, and to test them with several bodies of data

6. Summary

The findings from this extension of my earlier paper on "The Theory of Marriage" (1973) include

a) An explanation of why persons who care for each other are more likely to marry each other than are otherwise similar persons who do not

---

[35] See the discussion in section 3 of the Appendix

[36] The distinction between general and specific investment is well known, and can be found in Becker (1964, chap 11) Children, for example, would be a specific investment if the pleasure received by a parent were smaller when the parent was (permanently) separated from the children

[37] If they have relatively large differences because they were less efficient searchers, they may be less likely to separate

This in turn provides a justification for assuming that each family acts as if it maximizes a single utility function

b) An explanation of why polygyny, when permitted, has been more common among successful men and, more generally, why inequality among men and differences in the number of men and women have been important in determining the incidence of polygyny

c) An analysis of the relation between natural selection over time and assortive mating, which is relevant, among other things, for understanding the persistence over several generations of differences in incomes between different families

d) An analysis of which marriages are more likely to terminate in separation and divorce, and of how the assortive mating of those remarrying differs from the assortive mating in their first marriages

The discussion in this paper is mainly a series of preliminary reports on more extensive studies in progress The fuller studies will permit readers to gain a more accurate assessment of the value of our economic approach in understanding marital patterns

Mathematical Appendix

1 Formally, M (or F) maximizes his utility function

$$U_m = U_m(Z_m, Z_f) \tag{A1}$$

subject to the constraints

$$\left.\begin{aligned} Z_m^0 - C_m &= Z_m \\ Z_f^0 + C_m &= Z_f \\ C_m \geq 0 & \end{aligned}\right\}, \tag{A2}$$

where Z_m^0 and Z_f^0 are the market allocations of output to M and F, and C_m is the amount transferred by M to F If $C_m > 0$, these constraints can be reduced to a single income constraint by substitution from the Z_f into the Z_m equation

$$m_{mf} = Z_{mf} = Z_f^0 + Z_m^0 = Z_m + Z_f, \tag{A3}$$

where Z_{mf} is the output produced by M and F, and m_{mf} is M's income. Maximization of U_m subject to this single income constraint gives

$$\frac{\partial U_m}{\partial Z_m} = \frac{\partial U_m}{\partial Z_f} \tag{A4}$$

If $C_m = 0$, U_m is maximized subject to the two constraints $Z_m^0 = Z_m$ and $Z_f^0 = Z_f$ The equilibrium conditions are $\partial U_m/\partial Z_m = \lambda_m$, $\partial U_m/\partial Z_f = \mu_m$, where λ_m and μ_m are the marginal utilities of additional Z_m^0 and Z_f^0, respectively. The income of M would then be

$$m_{mf} = Z_m^0 + (\mu_m/\lambda_m)Z_f^0, \tag{A5}$$

where μ_m/λ_m is the "shadow" price of Z_f to M in terms of Z_m.

Since $\mu_m/\lambda_m < 1$ (otherwise $C_m > 0$),

$$Z_m^0 + \frac{\mu_m}{\lambda_m} Z_f^0 < Z_{mf} = Z_m^0 + Z_f^0. \tag{A6}$$

If $C_m > 0$, the "family" consisting of M and F would act as if it maximized the single "family" utility function U_m subject to the single family budget constraint given by (A3), even if F's utility function were quite different from U_m In effect, transfers between members eliminate the conflict between different members' utility functions.

2. Total income in a marriage between M and F is

$$m_{mf} + f_{mf} = I_{mf} = Z_{mf} + p_m Z^f_{mf} + p_f Z^m_{mf},$$

where I_{mf} is the total income in the marriage, Z^m_{mf} and Z^f_{mf} are the outputs allocated to M and F, Z_{mf} $(= Z^f_{mf} + Z^m_{mf})$ is total output, p_m is the shadow price to M of a unit of Z^f_m, and p_f is a shadow price to F of a unit of Z^m_{mf} Their incomes must be in the intervals

$$\begin{aligned} Z^m_{mf} + p_m Z^f_{mf} &= m_{mf} \le Z_{mf}, \\ Z^f_{mf} + p_f Z^m_{mf} &= f_{mf} \le Z_{mf} \end{aligned} \tag{A7}$$

If $p_m = p_f = 0$—no caring—m_{mf} and f_{mf} can be anywhere between 0 and Z_{mf} But if $p_m = p_f = 1$—mutual and full caring—then $m_{mf} = f_{mf} = Z_{mf}$ And, more generally, if p_m and $p_f > 0$, then

$$\begin{aligned} Z^m_{mf} &< m_{mf} \le Z_{mf} < I_{mf}, \\ Z^m_{mf} &< f_{mf} \le Z_{mf} < I_{mf} \end{aligned} \tag{A8}$$

Consider the following matrix of total *incomes*

$$\begin{array}{ccc} & F_1 & F_2 \\ M_1 & 8 & 8 \\ & & (4, 4) \\ M_2 & 7 & 7 \\ & (3, 4) & \end{array} \tag{A9}$$

On the surface, both sortings are equally optimal, but this is not so if only M_1 and F_2 have a marriage with caring, say full and mutual, so that $m_{12} = f_{12} = 4$ [38] The sorting M_1F_2 and M_2F_1 is not as viable as the sorting M_1F_1 and M_2F_2 because income is more divisible between M_1 and F_1 than between M_1 and F_2 [39] For if, say, $m_{11} = 4\frac{1}{2}$, $f_{11} = 3\frac{1}{2}$, $m_{22} = 4\frac{1}{2}$, and $f_{22} = 2\frac{1}{2}$, no two persons have an incentive to change mates and marry each other [40] On the other hand, since $m_{12} = f_{12} = 4$, unless $m_{21} = 3$ and $f_{21} = 4$, either M_1 and F_1, or M_2 and F_2 would be better off by marrying each other If $m_{21} = 3$ and $f_{21} = 4$, M_1 and F_1, and M_2 and F_2 could be just as well off by marrying each other Therefore, this sorting is not as viable as the sorting that does not have any marriages with caring

3 Assume that the gain from marriage of a particular person M is positively related to the expected values of two traits of his mate, as in $m = g(A_1, A_2)$, with

---

[38] The *output* between M_1 and F_2 also equals four, half that between M_1 and F_1

[39] Or, put differently, the output between M_1 and F_1 exceeds that between M_1 and F_2

[40] F_2 would prefer to marry M_1, but could not induce M_1 to do so because m_{12} cannot exceed four, the output produced by M_1 and F_2 (see eq [A7]), which is less than $m_{11} = 4\frac{1}{2}$.

$\partial g/\partial A_i = g_i > 0$, $i = 1, 2$. If the marginal costs of search were c_1 and c_2 for A_1 and A_2, respectively, equilibrium requires that

$$\frac{g_1}{g_2} = \frac{c_1}{c_2}. \tag{A10}$$

The lower c_1 is relative to c_2, the higher generally would be the equilibrium value of A_1 relative to A_2, since convexity of the isogain curves is a necessary condition for an internal maximum.

If g_1 and g_2 were invariant when search costs changed to all participants in the marriage market, not an innocuous assumption, then A_1^{max} and A_2^{max} would be the equilibrium values of A_1 and A_2 to M when everyone had perfect information about all traits A reduction in the cost of searching A_1, therefore, would move the equilibrium value of A_1 to M closer to A_1^{max}, its value with perfect information

Comment: The Economics of Nonmonetary Variables

William J. Goode

Columbia University

In the history of science researchers have often borrowed theories, analogies, or metaphors from other fields, usually the better-developed ones; in economic terms, they invested their human capital by acquiring new and presumably more-advanced intellectual tools. The most conspicuous borrowing in nineteenth-century social science was the unfortunately imaginary set of the developmental sequences of societies, worked out by anthropology and sociology on the basis of findings from biological evolution. It is less often that scientists in a relatively developed field become restive with its constraints and invade another with the aspiration that their more powerful technical and theoretical tools will solve problems with which the less-developed field has not adequately coped. (This is a challenge I have sometimes hurled at physicists who believe that the findings of sociology are simple-minded.) Over the past two decades, physicists tried this successfully in their contributions to molecular biology, and now economists have been expanding their world by attempting to analyze problems usually brooded over by sociologists, social psychologists, anthropologists, and political scientists.

In this case, interestingly enough, they will encounter colleagues in these fields who have themselves been moving toward economic or quasi-economic analysis of the same phenomena, though to be sure they have not attempted many social analyses of purely market processes. Especially in social psychology and sociology, a small group of theorists have for nearly two decades been working out exchange and allocational problems with the aid of economic ideas.

They have been hampered by their lack of mastery over the tools of economics, their failure to use economics explicitly, and their unwillingness—thus shouldering a burden economists of the past did not wish to carry—to do much theorizing without the facts. In my own case, I have been trying to understand social-control systems, of which monetary controls form one set, by considering how prestige, force and force threat, and love or affection are accumulated, allocated, or lost. Even in my monograph on divorce, written two decades ago, I was the first both to demonstrate rather fully the inverse relationship between class position and divorce rates (a prediction now made by Becker) and to give an essentially economic explanation for it (see Goode 1956, 1962).

The differing styles of economics and sociology prevent me from making an adequate, brief analysis of Becker's paper. I perceive many tautologies in the paper, but I know that disturbs an economist less than a sociologist. Where economists may be content with certain types of summary indexes, I want to see a large cross-table of percentages. Often I must respond by saying, in answer to an elegant set of mathematical formulas: "It is a beautiful flight but it is not reliable for transportation", that is, it simply is not true—though that may seem to an economist only a crude answer. An example of this would be Becker's formulation that late marriage and the difficulty of divorce will be correlated. India, for example, has an early age of marriage but very difficult divorce; in the West the main line of division is between Catholic and non-Catholic countries, and in general Catholic countries have almost no divorce but a slightly earlier age at marriage (except for Ireland) than other Western countries, and so on. Of course, to an economist, that may seem at best an unimaginative answer.

In any event, though I am nervous about this invasion, I welcome it. In these few pages I shall look at only a few minor points in Becker's paper to illustrate a somewhat general problem that is often encountered in some explorations by economists.

A major source of weakness in this bold foray into intellectual fields such as sociology and political science that have been trying for generations to create an autonomous noneconomic body of theory is Becker's failure to be daring enough in a critical question, that is, whether any nonmonetary variables actually enter the calculations. In fact, as we see from the very first stages (Becker 1973, p. 817), both the market and nonmarket variables are aggregated to explain S, which is the "full income, the maximum money attainable if the w_j [i.e., wage rates, of the jth member] are constants."

Although I believe this monetary emphasis may ultimately be a weakness, since the noncontractual structure of exchanges in those sectors may create different behaviors than in the monetary sector,[1] we are left with no assurance that the broadened economic formulation will be adequate, since in fact it has not been fully built into the equations or tested. That is, the nonmonetary variables have not been given due weight, but have been monetized

---

[1] This matter is analyzed at length in my *The Celebration of Heroes: Prestige as a Control System* (forthcoming). Meanwhile, see my *Explorations in Social Theory* (1973).

That monetary equivalent, so tempting to the economist because of his past training, can create both obstacles and factual errors in the analysis Perhaps I should add a theoretical point, that since the income figure is purely hypothetical or imaginary, it has no greater degree of reality than an apparently subjective factor such as prestige

To take a minor example, education *is* worth something on both the market and nonmarket exchange systems, and for some purposes a high education can be given an equivalent monetary value, but for a man with *little* education, a wife with very *much* education is *not* worth as much as one would suppose from her money value on the larger market She is a less-fit wife for him, by nonmonetary calculations—which would still be economic—and he would be supported in this low evaluation by his social circle

Let us consider at length a more elaborate instance of the problems this failure creates, specifically the lines of reasoning as to whether and why polygyny occurred and why it has usually been the upper-strata men who could or did take advantage of this opportunity The case is instructive, for the focus on total monetary income leads to factual errors and obscurities, but a general focus on nonmarket economic analysis would clarify some of the facts we do know and leave one question unsolved but at least not obscure

Becker reasons—correctly, I think—that the total monetary output of monogamy would be higher than that of either polyandry or polygyny Here the facts are more powerful than he supposes, for whether or not multiple marriage was permitted or encouraged, it has not been common in any society, except under very special conditions For most men, in all societies over time, polygyny was a statistically unlikely delight

Obviously, a very low ratio of men to women—caused by successful conquests or by high death rates in war—might permit polygyny for a while, but such conditions are not likely to continue for long Late marriage for men and early marriage for women raise the chances somewhat, but high mortality among women keeps the total number of woman-years available to each man rather limited As a consequence, polygyny has not been general even where it is approved

A purely market explanation may yield this aggregate result even if, as I believe, the reasoning in its favor (see his section 3, above) is tautological That is, for most men and women the net market payoff is higher if they pair together rather than make other types of arrangements However, this reasoning fails at the next step, the explanation for the unequal distribution of women among men even where polygyny is permitted

At this next step it is necessary to give due weight to strictly nonmonetary factors. First, Becker argues (see his section 3, above) that the "total output over all marriages could be greater if a second wife to an able man added more to output than she would add as a first wife to a less able one." Again leaving aside the tautological reasoning, this hypothetical result, an increase in total output over all marriages, is essentially empty as a motivating personal or social force in marital decisions.

As in many family decisions, especially those concerning fertility, the individual who decides does not ordinarily concern himself very much with his or her effects on the total output, or society as a whole. Thus, whether total output over all marriages is greater does not motivate people to enter polygyny. If Becker disclaims this as a motive, it is nevertheless obvious from the context that it is viewed as a partial explanation for the distribution of wives in a society.

Nevertheless, if we ask only about higher or lower monetary output of individual marriages, under Becker's assumptions we could expect that able men would try to get a second wife, and their chances of success would be greater than those of less able men

If nothing else were involved, we could be satisfied that the unequal distribution of wives among the total number of husbands is now explained, at least from the man's side (from the woman's side, the matter is less clear, since under polygyny the husband takes a much larger share than the wife does). Indeed we might then take a step further and predict which economic strata of men would be able to reap this advantage, and even how much they would pay in order to make this investment in a wife, as compared with alternative investments in other types of capital goods Where, for example, as in much of Central Africa, women may earn money by engaging in small-scale trading, or by tilling their own gardens for the family, we might even be able to put real monetary values on this kind of investment. Moreover, we might even give a better explanation for polyandry than Becker essays, since for the most part such systems occur where the productivity of a given man is low and he has little to invest in a wife, so that a set of men (often brothers) pool their assets in order to enjoy the benefits of one wife.[2]

However, reasoning from market or monetary total output, whether of the marriage or of all marriages, would miss the importance of women as goods, of women as producers of children, of women as links in purchasing political power, of women as prestige commodities, for monetary reasoning would *not* predict that women would be added where their monetary market output drops or almost stops upon entering marriage. For the stubborn fact remains that multiple marriage is most common among the stratum that gets the least monetary economic output from wives, that is, the most powerful and rich, who are most likely to wall them up and use them as pets, display them as objects that prove the man's high position, use them in power linkages, or even allow them to become parasites who destroy the family wealth by their extravagance Indeed, it seems reasonable to assert that the higher the class position, the greater the likelihood of some polygynous arrangements, but equally the greater the likelihood that wives represent net losses, except in purely nonmarket goods Like some other monetary goods, they are in part "bought in special markets", they are also like such goods in another respect, for instance, furs, polo ponies, Rolls Royces, in that they can be put to other practical uses, and in revolutions sometimes are

---

[2] In polyandry these are often brothers, perhaps because this causes less strife; and this alleviates somewhat the problem of producing one's *own* children, which Becker speaks of as a goal in marriage. In any event, whether or not the husbands are brothers, there are usually rules to decide who is sociologically defined as father of a given child.

Thus, I am asserting that the reasoning of Becker's paper sometimes goes awry not because it uses economic reasoning in nonmarket areas, but because it fails to take note of the powerful nonmarket variables, whose effects may run counter to those of income variables, however these latter are aggregated Nor am I as yet sure that market prices or decisions "prove" the total output is greater, whether of general nonmonetary or monetary goods, under these arrangements As Frederick Knight said more than half a century ago, all suppliers are partially monopolies, and we know that these markets are rigged

In a parallel fashion, the reasoning fails in its attempt to explain the decline of polygyny, that is, the system, in our time and to some extent (although the facts are less clear) over a longer period of time—less clear because we suppose that never in the history of the recent world have most nations been polygynous except under rare conditions

After flirting with a biological explanation, Becker involves himself in a complex refutation of the claim that women would be better off economically if polygyny were abolished (see his section 3, above) But his argument, and those he tries to refute, are essentially irrelevant, because all of them still seem to refer to the total market output of the household It is much more likely that nonmonetary variables shape this decline in polygyny

Here, as so often in such ventures into why social structures assume their sometimes peculiar shape, we have to ask, not merely about the nonmarket variables, but *who decides* who organizes the market In spite of Becker's graceful reference (n 27) to Shaw's clever comment, the fact is that we have no evidence from any time or place that *women* ever created a polygynous system It was *men* who established them, men whose command of force and force threat, prestige, and wealth was paramount in all the societies we know Doubtless, some women did better economically under such systems, as some women in the United States would if they joined with other women to marry one rich and powerful man, but the system as a whole was not created by women, and wherever new laws, social movements, revolutionary ideologies, industrialization, urbanization, or religion permit them a free market choice they move away from polygyny[3]

I shall not presume, especially in the midst of our contemporary turmoil about who has the right to report what women feel, to know precisely why women object to this system, although I think Becker's earlier, excellent analysis of caring and sharing are relevant. From the known fact that the birth rate under polygyny is less than under monogamy, and from the rules about equal sexual rights in polygynous systems, we must infer that women do not get as much tenderness, affection, and sexual enjoyment in those systems—again, we can use economic theory to reason about these variables, but they are not monetary. Men are in a much more influential position in such systems, and do not have to share as much of the physical output of the family, as indeed they also monopolize much of the prestige-esteem output of the family. They have the power to impose their own prices, in such systems Thus, when external constraints, including money-market constraints, permit, women decline to enter such unions, even when they are still permitted.

I suggested earlier that after this critical commentary an important puzzle would still remain, although it is one that Becker has really not perceived clearly It is why some social *systems* are polygamous, and others are not An explanation of a *system* may not be adequate as an explanation of how some individuals act within it The possibility that some men will benefit from taking more than one wife does not explain why polygyny occurs here rather than there—and it especially fails if we remember that even in most such systems the typical marriage is a pairing, one man to one woman Why are such systems approved and viewed as an *ideal,* though few can attain it?

I do not think that any current explanations work well, though doubtless when one does it will be some sort of nonmarket economic analysis That explanation would have to confront the fact that Western societies, including Latin America, of course, have not been polygynous in any historical period—and of course I do not mean only in the statistical or distributional sense, but in the normative sense, that polygyny was never a socially approved pattern of marriage. Both Japan and China have been mainly monogamous, though in both cases wealthy or powerful men could purchase additional concubines, and in China some of these came close to being recognized as "secondary wives" with specific rights India, similarly, has been mainly monogamous, though again with a much more open acceptance of concubines. Perhaps most societies outside those great civilizations—but remember they encompass most of the world's population—have been polygamous in the specific but narrow sense that some form of polygamy was normative, something to which men and families aspired

---

[3] See in this connection the relevant sections of my *World Revolution and Family Pattern* (1963), esp. chap. 4 on Sub-Saharan Africa.

Whatever those explanatory variables will turn out to be when we locate them, I rather suppose they will not be monetary, and—more important—their impact will be seen primarily among the people who dominated those sociopolitical systems and those marriage markets, that is, the upper-strata men. Upper-strata men created a set of constraints that defined the marriage and family market so as to yield great monetary and nonmonetary advantages to them

Few good theories in the history of science "fitted the facts" very well, and the "facts" at any given time were partly wrong, so that each explorer must follow his or her temperament rather than a rule which would ground new thinking only on the facts or only on free speculation I do suspect, however, that a deeper sense of the crude regularities would not hinder these elegant flights, but might pose still more interesting puzzles for the imaginative economist More fundamentally, I am at least partly persuaded by the work before us that in several fields we are moving toward a single structure of hypotheses about social behavior, hypotheses supported by research in several different fields under different labels, and that two decades hence we may discover we have created a rather impressive body of social science—if not "unified," at least mutually confirmatory Becker's explorations are an important step in that progress

The United States Marriage Market

Alan Freiden

Virginia Polytechnic Institute and State University

I. Introduction

The empirical implementation of theoretical contributions in the social sciences usually rests on variables defined and measured by persons other than the theorist Such is the case with the theory of marriage For the United States, the only relevant demographic quantity measured with accuracy is two-person, male-female legal marriage. Thus, my decision to study legal marriage was imposed by the Bureau of the Census and the inherent difficulties of measuring other forms of communal living This does not seem to be restrictive for the United States, where consensual marriage is often a prelude to legal marriage, but it may be important in countries such as those in Latin America, where legal and other social institutions (i e , divorce laws) differ [1]

That a model of marriage based on rational choice has substantial explanatory power is demonstrated in this study Three factors—the ratio of the sexes, the potential returns to marriage, and the cost of divorce—explain many of the areal differences in marital behavior in the United States. It is also evident that the marriage market has important indirect effects on the reproduction rate

II. The Marriage Market

People decide to marry when they expect to enjoy some flow of "real" income, denoted by Z, which is greater than that which they could receive

---

I thank Gary Becker, H. Gregg Lewis, Marc Nerlove, Margaret Reid, T W Schultz, and T. Dudley Wallace for their many helpful suggestions, not all of which have been incorporated in this paper I, of course, am solely responsible for any errors. The work for this paper was financed in part by grants to the University of Chicago from the National Institutes of Mental Health and the Rockefeller Foundation.

[1] For an analysis of the three forms of marriage recognized in Brazil, see Kogut (1972) Cheung (1972) has described the structure of property rights within marriage in China

if single.[2] The Z represents an abstract commodity, produced within households, which is the true source of personal satisfaction Let

$$Z^{mf} = Z^{mf}_{m} + Z^{mf}_{f}{}^{3}$$

represent the yield to marriage, where the right-hand side is the distribution of this "real" income between the married partners With Z^{mf} taken as given, this distribution of married real income is determined by the supply of husbands and wives to the marriage market The qualitative results derived holding Z^{mf} constant are not altered when couples are sorted in a manner which optimizes the total output of Z within the community

If unmarried, the ith male would obtain the real income Z_{mi} and the ith single female would obtain Z_{fi} Either would prefer marriage if and only if

$$Z_{mi} \leq Z^{mf}_{m}$$

or

$$Z_{fi} \leq Z^{mf}_{f}.$$

Since no male will accept less than his single income, the first male desiring marriage will be forthcoming at some level of married income, Z^{mf}_{m}, which is equal to the minimum single income, Z_{mi}, taken over all males By ordering males with respect to their single incomes, the supply curve of males in the marriage market is equivalent to the cumulative distribution of males with respect to the Z_{mi} In figure 1, this distribution has been drawn as if the maximum single income were undefined, that is, the supply curve asymptotically approaches the boundary set by the total number of males (which initially equals the number of females) The qualitative results derived here do not depend on this particular distribution Similarly, the supply curve of females is the cumulative distribution of the Z_{fi} Now, since

$$Z^{mf}_{f} = Z^{mf} - Z^{mf}_{m},$$

the supply of females is a negatively inclined function of the male marital income, Z^{mf}_{m}, and can be viewed as the derived demand for husbands by females [4] No women can be found who would offer more than Z^{mf} less the minimum of Z_{fi} for a husband, so the supply-of-potential-wives curve begins at

$$Z^{mf} - Z_{fi}\,(\text{min})$$

and falls to the right.

---

[2] This section is based on Becker (1973) For another application of the general model see Keeley (1973)

[3] The superscript mf denotes a married couple. The subscripts m and f refer to single males and females.

[4] The analysis is, of course, perfectly symmetrical for the derived demand for wives by males

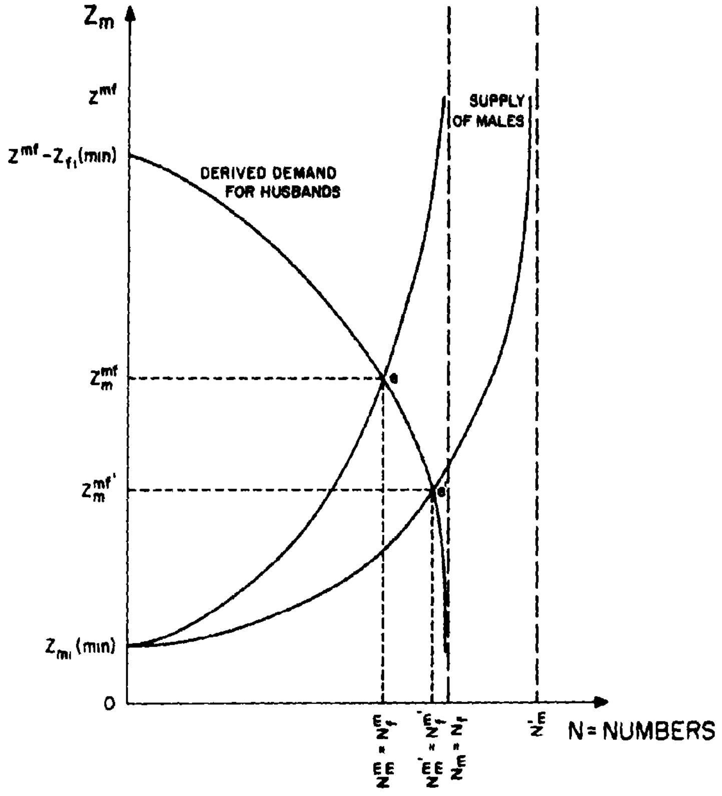


Fig 1 —The marriage market

The intersection, e, of the supply curves in figure 1 gives the number of men and women marrying as well as the distribution of income within marriage The proportion of females married will be N^{m}_{f}/N_f, and their real income will be $Z^{m}_{f} = Z^{mf} - Z^{mf}_{m}$

Now, consider an increase in the number of men to N_m'. This will shift the equilibrium to e', raising the real income of married women to

$$Z^{m}_{f}{}' = Z^{mf} - Z^{mf'}_{m} > Z^{m}_{f} = Z^{mf} - Z^{mf}_{m}$$

This increase is not directly observable, but the coincident increase in the number of females married, from N^{m}_{f} to $N^{m}_{f}{}'$, will raise the proportion of women married to

$$\frac{N^{mf'}}{N_f} > \frac{N^{mf}}{N_f}.$$

Thus, the model predicts that the proportion of women married should be positively related to the ratio of men to women

To derive further implications, the nature of the household production process generating Z^{mf} (and the single real incomes) must be made explicit Let

$$Z = g(X, T_m, T_f)$$

represent the household production function for the flows of real income previously denoted by Z^{mf}, Z_m, and Z_f The X represents goods and services purchased in the market, T_m is male time, and T_f, female time The g is then the household production function which summarizes the technology available for converting market goods and time into the actual source of utility, namely, Z As increasing returns to scale would always result in a gain to marriage regardless of the number or sex of the participants, g will be a function such that economies of scale are ruled out [5]

A single male would maximize his production of Z subject to the constraint that $T_f = 0$, that is, Max $Z = g(X, T_m, T_f)$, subject to (1) $T_m + L_m = T$, L_m = market work, (2) $PX = W_m L_m + V_m$, P = index of prices, W_m = wage rate, V_m = nonmarket income, and (3) $T_f = 0$ The maximized value of Z is Z_m Similarly, a single female would Max $Z = g(X, T_m, T_f)$, subject to (1) $T_f + L_f = T$, (2) $PX = W_f L_f + V_f$, and (3) $T_m = 0$ The maximized value of Z is Z_f In either case, one person must substitute his own time to make up for the absence of the other's A married couple would maximize Z without the third constraint Max $Z = g(X, T_m, T_f)$, subject to (1) $T_m + L_m = T$, $T_f + L_f = T$, and (2) $PX = W_m L_m + W_f L_f + V_m + V_f$ The maximized value of Z is Z^{mf} The gain to marriage is then

$$G = Z^{mf} - (Z_m + Z_f)$$

A more convenient form of this expression is obtained as follows If $Z = g(X, T_m, T_f)$, where g is a function of factor inputs, then there exists a function $c = c(P, W_m, W_f)$, a function of factor prices, which gives the minimum average cost of producing units of Z in the household (this is the cost dual of the household production function) For married couples this function is $c^{mf}(P, W_m, W_f)$, for single males it is $c_m(P, W_m)$, and so on The maximized value of Z is then total resources available divided by this average cost of production Resources available, called full income, can be obtained by combining the time and goods constraints For a married couple these constraints are (1) $T_m + L_m = T$, $T_f + L_f = T$, and (2) $PX = W_m L_m + W_f L_f + V_m + V_f$

---

[5] Note that a gain to marriage is not ruled out by decreasing returns to scale

Let

$$L_m = T - T_m \quad \text{and} \quad L_f = T - T_f,$$

then

$$PX = W_m(T - T_m) + W_f(T - T_f) + V_m + V_f, \text{ or}$$
$$PX + W_mT_m + W_fT_f = (W_m + W_f)T + V_m + V_f.$$

The left-hand side is the total monetary equivalent of time and market-goods expenditures, while the right-hand side is full income. Thus,

$$G = \frac{(W_m + W_f)T + V_m + V_f}{c^{mf}(P, W_m, W_f)} - \left[\frac{W_mT + V_m}{c_m(P, W_m)} + \frac{W_fT + V_f}{c_f(P, W_f)}\right]$$

By totally differentiating G, further implications of the model may be derived [6]

First, consider the wage rates If

$$W_m = W_f,$$

$$MP_m = \frac{\partial Z}{\partial T_m} = MP_f = \frac{\partial Z}{\partial T_f}, \quad \text{and}$$

$$V_m = V_f, \quad \text{then}$$

$$Z_m + Z_f = 2Z = Z^{mf}.$$

This means that if male and female time are perfect substitutes in both household and market production, there is no gain from marriage However, in all other cases Z^{mf} will exceed the combined single incomes Note that the marginal products must be equal throughout the range of inputs and not just at a point In particular, the combined gain would be greater the larger the difference between W_m and W_f If, as is usually the case, $W_m > W_f$, then the husband will specialize in market production while the wife concentrates her time in the home Therefore, the gains from marriage and the proportion of females married should be positively related to the relative wage rate, W_m/W_f. However, the effect of equal proportionate increases in the male and female wage rates is ambiguous On the one hand, there is a positive income effect associated with the rise in full income. But on the other hand, there is a negative cost-of-production effect due to the elasticity of the cost dual of the household production function with respect to the wage rates If this elasticity is sufficiently large (in absolute value), the net effect of equiproportionate increases in male and female wages may be negative. Note that equal percentage increases in the nonlabor incomes V_m and V_f do not generate a negative cost-of-production effect, so the gains to marriage are expected to rise with such a change

---

[6] This approach is equivalent to a comparative statics analysis of the market-supply curves, since an increase in the gain to marriage shifts both curves to the right in figure 1

The manner in which education affects the household production function is not obvious, so the prediction of this factor's effects on the gains to marriage is unclear For example, consider the following two ways of looking at education· first as factor-neutral technological change and then as factor-augmenting In the first case, education uniformly reduces the amount of each input required to produce units of Z, that is,

$$Z^{mf} = \frac{(W_m + W_f)T + V_m + V_f}{A_m(E_m)A_f(E_f)c^{mf}(P, W_m, W_f)}, \text{ and so on,}$$

where E stands for education and the A's are functions whose first derivatives are negative It can be shown that the gain to marriage is positively related to increases in education in this case If treated as factor-augmenting technical change, education improves the efficiency of the time input independently, that is,

$$Z^{mf} = \frac{(W_m + W_f)T + V_m + V_f}{c^{mf}[P, B_m(E_m)W_m, B_f(E_f)W_f]},$$

where B_m and B_f are functions similar to the A's Special conditions must hold to obtain the same results as above If education raises the nonmarket productivity relative to the market productivity of the mate specializing in household production, or the relative market productivity of the market specialist, then it increases the gains to marriage However, either of these possibilities may prove difficult or impossible to distinguish empirically.

Until this point, an implicit assumption was made that perfect information and certainty held for persons in the marriage market Clearly, this is not the case The high divorce rate of persons married 2 years or less is evidence that uncertainty plays an important role This divorce behavior may be interpreted as the correction of mistakes Then a lowering of the cost of correcting a mistake (the difficulty of divorce) should induce people to marry more readily The effect of the cost of divorce on the stock of marriages, however, is uncertain since a fall in the cost of divorce also encourages the dissolution of existing marriages The former is likely to dominate the behavior of the young, while the latter is more relevant for older couples

III. Empirical Implementation

Three sets of data—state, county, and Standard Metropolitan Statistical Area (SMSA) aggregates—are used to evaluate the model's predictions Since no single set of data is completely satisfactory, this test rests on evidence accumulated from several sources The most complete data are at the state level, but the variables are highly collinear. Heterogeneity among the observations is increased by using a random sample of United States counties, but the federal census compilations are insufficiently detailed, especially with regard to age-specific and female wage data.

The method used is to estimate an ordinary least-squares regression with a measure of the proportion of females married (FEMMAR) as a dependent variable and the relative wage (RELWG), the ratio of the sexes (SXRT), the cost of divorce (DCOST), and a measure of income as independent variables.[7] All observations are weighted by the appropriate number of females. For the states and SMSAs, equations are estimated for each of the 5-year age cohorts between ages 15 and 39. The actual variables used are described below and listed in table 1.

1. FEMMAR. The proportion of females married is the number of women married, with spouse present, divided by the total number of women. There is a small error due to the exclusion of married women having a spouse absent for reasons other than separation, but this is insignificant. More important, however, is the treatment of widows. On the average, males die earlier than females (especially married males), so there will be some widows who should be considered "involuntarily" single. Investigating a revised figure for females married, one that included widows, showed little difference in the results, so this factor is ignored.

2. RELWG. Three alternative measures of male and female wage rates are available. The best is the annual earnings of full-time wage and salary workers as estimated by the Social Security Administration. These figures represent the marginal (and here average) opportunity cost of time to a prospective labor-market participant. It is likely that these values approximate the permanent wage rates for workers by state. The relative wage is simply male earnings divided by female earnings. The mnemonic for the Social Security wage is RELWGSS. A second measure of the full-time wage is the United States Census figure for median earnings of persons working 50–52 weeks in the previous years. Since this is available for both states and SMSAs, direct comparisons between the results are feasible. The symbol for a measure of the male-female relative wage based on census data is RELWGCN. Finally, it is most unfortunate that neither of these measures is compiled at the county level. The only sex-specific earnings data are the median annual earnings of persons. This value is quite sensitive to labor-force participation rates, so it is a biased measure for the purpose of explaining marriage. Suppose the full-time relative wage is high, then the gains to marriage are highest with husbands

---

[7] Education cannot be investigated at this time. In aggregate data, the relationship between income and education is very close, so multicollinearity is a severe problem. Later, the 1970 Public Use Samples of Basic Records will be available from the Bureau of the Census for further empirical work.

TABLE 1
VARIABLES AND DATA SOURCES

Symbol	Description	Source
FEMMAR*.	Proportion of females married, spouse present	U S Bureau of the Census (1962, 1963*b*, tables 59 and 105)
RELWGCN	Male/female relative wage, persons working 50–52 weeks in 1959	U S Bureau of the Census (1963*b*, table 134)
RELWGSS	Male/female relative wage, full-time annual earnings	U S Department of Health, Education, and Welfare (1968*b*, table V-17)
RELWGAE	Male/female relative median annual earnings	U S Bureau of the Census (1963*b* tables 134 and 136)
SXRT*	Sex ratio, number of males/number of females	U S Bureau of the Census (1963*b*, table 59)
DCOST	Index of ease of divorce	Table 2 (below)
CATH	Proportion of population Catholic	National Council of Churches (1956, table 4)
HSEVAL	Median value of owner-occupied housing	U S Bureau of the Census (1962, 1963*b*, table 20)
RENT	Median gross monthly rental	U S Bureau of the Census (1962, 1963*a*)
BIRRATE	Births per 1,000 women	U S Bureau of the Census (1962)
AGRI	Proportion of the labor force employed in agriculture	U S Bureau of the Census (1962, 1963*b*, tables 82 and 128)
NONWT	Proportion of the population non-white	U S Bureau of the Census (1962)
FEMEARN	Median annual earnings of females	U S Bureau of the Census (1963*b*, tables 86 and 136)
FEMSCH	Median school years completed, females, 25 years and older	U S Bureau of the Census (1963*b*, table 83)
MALSCH	Median school years completed, males, 25 years and older	U S Bureau of the Census (1963*b*, table 83)
AGE	Index of age distribution based on national age-specific rates	†

---

* Age-specific data

†
$$\mathrm{AGE} = \frac{\sum_{k=1}^{15} w(k) * NF(k)}{\sum_{k=1}^{20} NF(k)},$$

where $NF(k)$ is the number of females in the kth age group $k = 1, 2, \quad , 20$ for ages 0–4, 5–9, , 95–99,

k	w(FEMMAR)	w(BIRRATE)
1	0	0
2	0	0
3	0 2	0 8
4	15 7	89 1
5	69 5	258 1
6	86 2	197 4
7	88 7	112 7
8	88 2	56 2
9	85 9	0 9
10	82 5	0
11	77 0	0
12	69 9	0
13	61 4	0
14	51 6	0
15	65 0	0

specializing in market labor and wives in household production. This implies a lower female labor-force participation rate and lower median annual female earnings Therefore, if the model is correct, the observed annual earnings of females will be negatively correlated with the proportion of females married With this caveat, a relative wage based on annual earnings (RELWGAE) is used for the counties.

The preceding analysis brings to light an interesting question about the interpretation of observed wage differentials and also their relation to education If formal schooling primarily acts to raise market-earning potential, then the measured effects of education, holding the market wage constant, must be picking up something else, that is, on-the-job training Suppose husbands and wives specialize as predicted. For a time men are receiving on-the-job training which raises their market age-earnings profile above that which would hold in the absence of such training. At the same time, their wives would experience depreciation of their stocks of market-specific human capital (but an increase in the nonmarket component) which lowers their permanent market wage. Therefore, even the full-time wage is somewhat biased as a predictor of marital behavior [8] It is true that a "pseudo wage" estimated from the demand side of the market would be more appropriate, but this approach will not be followed at present

3 SXRT The ratio of the sexes is simply the number of males in each age group divided by the number of females For the counties it is the number of males 15 years and older divided by the equivalent number of females. A minor concern with this definition is that males tend to marry females a few years younger than themselves Redefining the sex ratio to take account of this does not affect the results appreciably

4 DCOST The measure of the cost of divorce is not a quantitative variable. It is an index of the ease of divorce legislation as compiled through questionnaires and other data in 1959 [9] However, an attempt was made to achieve cardinality, so the actual values used are probably at least highly correlated with the true values Note that it is an index of the ease or permissiveness of the law and represents the inverse of the cost of divorce Table 2 lists the values used Religion may also be a relevant psychic cost of divorce for Catholic couples, so a set of regressions including the percentage of the population Catholic (CATH) in place of DCOST is estimated for the states

5 INCOME. The appropriate income measure is full family income, including labor and nonlabor components. The census value, median family income, is subject to the same measurement problems mentioned above for the female wage and is, therefore, unacceptable. An alternative,

---

[8] This may be added to Gronau's list of applications (1972) in which observed female wages are biased predictors of behavior

[9] See Rheinstein (1972) and Broel-Plateris (1961) for details

TABLE 2
CLASSIFICATION OF STATES BY THE EASINESS OF THEIR DIVORCE LAWS, 1959

Category and State	Easiness Score
Very permissive:	
Nevada	95
Florida	90
Arkansas	85
Idaho	81
New Mexico	77
Alabama	76
California	74
Wyoming	72
Permissive:	
Kentucky	69
Georgia	67
Arizona	66
Oklahoma	64
Utah	63
Washington	62
Mississippi	60
South Dakota	59
Oregon	59
Kansas	58
Indiana	57
Medium	
Montana	55
Tennessee	54
Colorado	53
Missouri	52
Texas	50
Iowa	49
Medium	
Nebraska	49
Ohio	48
North Dakota	46
Delaware	45
Vermont	44
Strict	
Maryland	42
Minnesota	41
West Virginia	40
Maine	39
Illinois	37
Louisiana	36
Virginia	34
Wisconsin	33
New Hampshire	32
Michigan	29
Dist of Columbia	28
Very strict	
North Carolina	25
Pennsylvania	22
Rhode Island	19
Connecticut	16
New Jersey	13
Massachusetts	9
New York	4
South Carolina	1

SOURCE.—Broel-Platers 1961, table 12, p. 149

on the expenditure rather than the receipt side of family accounts, has been suggested (Margaret Reid 1962) This is the median value of owner-occupied housing (HSEVAL) Since a large proportion of urban dwelling units are not occupied by owners, the median monthly gross rental (RENT) will also be used for the SMSAs. There are two reasons for using these measures. First, the observed positive income elasticity of demand for housing guarantees that a family's housing expenditure is positively correlated with income so HSEVAL and RENT will be unbiased surrogate measures. Second, these expenditures are not sensitive to the transitory component of income.[10]

IV. Results

There is a problem interpreting the regression results reported below The FEMMAR is a limited dependent variable, that is, a proportion bounded by zero and one (0–100 in percentages). This complicates the evaluation of elasticities but not the estimation technique, since no estimated values of the dependent variable violated these boundary constraints[11] Consider a given percentage change in a bounded variable. If the variable is already near its upper bound, then a small change is also a large change relative to the distance to the upper bound. For example, suppose FEMMAR is 90. A 1 percent increase to 90.9 is the same as a 9 percent decrease in the distance between the initial value and the upper bound Therefore, elasticities for the older age groups, those with an average proportion of married women greater than 80 percent, should be interpreted with this in mind

---

[10] There may be a bias due to a price component.

Tables 3 and 4 are the ordinary least-squares results from the state and SMSA data, respectively Two alternative measures of relative wages are shown for the states and two of income for the SMSAs Results for one set of equations from each data source are summarized in elasticity form in table 5 The explanatory power of the four independent variables is quite good, with $R^2 > .5$ in all but one case and with the vector of estimated coefficients significantly different from zero at the 5 percent level of confidence in every case. However, there are a number of parameter estimates with large standard errors; this is not unusual, given the level of aggregation employed here Two approaches to this problem will be described later

The data support the hypothesis that the proportion of females married is positively related to the ratio of the sexes Only one coefficient is not significantly greater than zero at the 5 percent level[12] Evidence on the effect of the cost of divorce is that the incidence of marriage rises as divorce becomes easier Table 6 contains the regression results for a replication of one set of the state equations using the proportion of the population Catholic (CATH) in place of DCOST Here a negative sign on CATH is equivalent to a positive one on DCOST, so the coefficients (all significant) are consistent with theory.[13] Note that for all five sets of regressions (state, SMSA, and state with CATH) the elasticity of marriage with respect to the cost of divorce declines monotonically with age The decline is greater than that which is generated by the effect discussed at the beginning of this section. This indicates that of the two effects mentioned earlier, encouraging new marriages and lowering the cost of dissolving existing ones, the former is more relevant for young couples The influence of income is puzzling Where significant, the coefficients of both RENT and HSEVAL are negative. This contradicts Silver's finding (1965) that the flow of new marriages (i e , the marriage rate) conforms to the business cycle. Of course, if the income elasticity of the divorce

---

[11] Elsewhere I have used the logit transformation in this context (1972)

[12] Of course, one coefficient out of 20 will randomly fail the 5 percent t-test when the null hypothesis is true

[13] Multicollinearity prohibits including CATH along with DCOST and HSEVAL in the same equation.

TABLE 3

REGRESSION OF FEMMAR, 1960 STATE DATA

INDEPENDENT VARIABLE	AGE 15–19	15–19	20–24	20–24	25–29	25–29	30–34	30–34	35–39	35–39
RELWGCN		−0.078		3.55		8.06		4.97		5.22
		(0.03)		(0.84)		(2.57)		(2.55)		(3.14)
RELWGSS	−2.76		4.59		7.15		3.32		3.74	
	(1.26)		(1.37)		(3.16)		(2.08)		(2.74)	
SXRT	22.28	28.24	13.10	11.01	43.59	52.26	60.55	68.19	62.72	71.72
	(2.72)	(3.66)	(1.85)	(1.55)	(3.68)	(3.99)	(6.92)	(8.05)	(7.63)	(9.05)
DCOST	0.099	0.094	0.137	0.140	0.026	0.018	0.016	0.012	0.009	0.004
	(7.68)	(7.29)	(6.67)	(6.72)	(1.50)	(1.01)	(1.50)	(1.16)	(0.99)	(0.47)
HSEVAL	−0.484	−0.489	−0.501	−0.449	−0.754	−0.702	−0.580	−0.565	−0.325	−0.293
	(4.54)	(4.36)	(3.02)	(2.60)	(5.10)	(4.65)	(5.90)	(5.87)	(4.21)	(3.85)
CONSTANT	−0.929	−11.52	46.45	49.96	36.58	27.20	28.41	18.66	22.99	12.16
	(0.08)	(1.06)	(4.20)	(4.01)	(3.23)	(1.90)	(3.71)	(2.12)	(3.17)	(1.49)
R^2	82	81	72	71	60	57	70	71	73	74
$F(4, 43)$	47.53	45.45	27.00	26.02	16.04	14.31	24.86	26.51	28.37	30.19

NOTE —t-values in parentheses

TABLE 4
Regression of FEMMAR, 1960 SMSA Data

Independent Variable	Age 15–19	15–19	20–24	20–24	25–29	25–29	30–34	30–34	35–44	35–44
RELWGCN	−0.640 (0.20)	0.041 (0.01)	2.19 (0.47)	9.31 (1.99)	6.04 (1.85)	10.98 (3.07)	1.65 (0.91)	4.43 (2.11)	1.48 (1.35)	2.70 (2.30)
SXRT	8.27 (2.95)	8.87 (3.18)	7.05 (1.98)	10.10 (2.76)	28.48 (3.23)	27.71 (2.61)	31.30 (4.74)	33.03 (3.41)	42.91 (9.49)	50.99 (9.15)
DCOST	0.125 (8.48)	0.137 (9.34)	0.183 (9.09)	0.191 (8.94)	0.084 (5.13)	0.100 (5.42)	0.014 (1.47)	0.025 (2.38)	−0.015 (2.54)	−0.015 (2.15)
HSEVAL	−0.395 (2.84)	.	−0.546 (2.80)		−0.584 (4.35)		−0.452 (5.31)	.	−0.242 (4.98)	. .
RENT	.	−0.121 (2.69)		0.005 (0.08)	.	−0.035 (0.67)	.	−0.072 (2.38)		−0.057 (2.88)
CONSTANT	7.55 (0.94)	8.81 (1.03)	56.49 (4.91)	33.21 (2.66)	51.95 (4.31)	38.00 (2.83)	60.52 (8.42)	52.67 (6.18)	46.90 (9.38)	37.97 (7.33)
R^2	.67	67	.70	.66	68	58	.56	.38	.73	.66
$F(4, 58)$	30.1	29.5	34.1	28.34	30.6	19.78	18.26	8.74	38.76	28.18

Note.—t-values in parentheses.

TABLE 5
SUMMARY OF REGRESSION RESULTS

INDEPENDENT VARIABLES	ELASTICITIES AT MEANS				
	Age				
	15–19	20–24	25–29	30–34	35–39
	State Data				
RELWGCN					
RELWGSS	−0.33	0.12	0.15	0.07	0.08
SXRT	1.48	0.20	0.52	0.70	0.73
DCOST	0.28	0.09	0.013	0.008	0.005
HSEVAL	−0.38	−0.09	−0.11	−0.08	−0.04
RENT					
	SMSA Data				
RELWGCN	−0.005	0.24	0.22	0.09	0.05
RELWGSS					
SXRT	0.63	0.15	0.33	0.38	0.59
DCOST	0.41	0.12	0.05	0.012	−0.007
HSEVAL					
RENT	−0.64	0.01	−0.03	−0.06	−0.05

TABLE 6
REGRESSION OF FEMMAR, 1960 STATE DATA

INDEPENDENT VARIABLE	AGE				
	15–19	20–24	25–29	30–34	35–39
RELWGSS	6.90	19.0	9.59	4.34	3.82
	(2.36)	(4.67)	(3.94)	(2.43)	(2.73)
SXRT	46.2	32.3	49.7	56.3	52.3
	(4.85)	(4.79)	(5.34)	(7.14)	(7.91)
CATH	−0.191	−0.303	−0.095	−0.065	−0.048
	(4.47)	(5.02)	(2.41)	(2.35)	(2.24)
HSEVAL	−0.110	0.096	−0.629	−0.497	−0.307
	(0.62)	(0.38)	(3.76)	(4.15)	(3.43)
CONSTANT	−38.8	5.31	28.5	32.8	36.1
	(2.89)	(0.42)	(2.93)	(4.59)	(5.97)
R^2	70	64	64	70	73
$F(4, 43)$	25.3	18.7	17.8	26.3	29.9

NOTE —*t*-values in parentheses.

rate exceeds that of the marriage rate, then the cross-sectional and time-series results are consistent

A strong theoretical result is that the potential gain to marriage and, therefore, the proportion of women married, is positively related to the sex-specific relative wage. This is confirmed for persons age 25 and older but not for the young Suppose that the benefits to marriage as perceived by teenagers differ from those of adults. Then the gains to specialization may not be relevant although other market factors (sex ratio and cost of divorce) are If so, then as these people mature, many of the "irrational" marriages will dissolve, leaving in the stock of married persons only those enjoying gains to specialization While ad hoc, this argument is plausible given the substantially higher divorce rate among couples married as teenagers [14] Also, suppose the costs of marriage for teenagers differ from those for adults. This is obviously the case, since the age of consent is 21 for males and 18 for females in most states. Then a measure of the cost of marriage as well as of the cost of divorce is necessary in the regressions for the group of persons ages 15–19 Unfortunately, no such measure is readily available, but the states with a substantial Catholic population seem to have the strictest marriage laws [15] If the results for persons ages 15–19 in table 6 are interpreted with CATH viewed as a measure both of the cost of divorce and the cost of marriage, the relative wage effect is substantiated

The evidence presented so far would have been more convincing had the estimates been more accurate Imprecise parameters may be improved through the application of additional sample information Here two techniques are employed to make better use of the data available First, it is obvious that the division of the population into 5-year age cohorts is arbitrary Therefore, one would expect the behavior of adjacent age groups to be similar For the regression model used here, the correlations of the residuals of adjacent age groups is very high There is information in the set of residual correlations which can be used to obtain more efficient estimates of the coefficients of all five equations through the seemingly unrelated regressions technique (Zellner 1962) The results for the state data (table 7) are as expected The standard errors of the coefficient estimates are improved, especially those for DCOST Since there are no qualitative changes in the coefficients, the discussion of the ordinary least-squares results need not be amended.

A second way to improve the information content of the sample is to expand the number of observations Roughly a tenfold increase in sample size occurs through use of a random sample of 530 United States counties [16] Unfortunately, no age-specific marriage data are compiled by the United States Bureau of the Census at this level, so the analysis must be restricted to a single equation aggregated over age To standardize the age structure of the county populations, an index (AGE) was constructed which has the value of FEMMAR if the proportion of women

---

[14] Education must be the next factor to be examined in this area. It is likely that something related to schooling, such as differential human capital investments or the more efficient college marriage market, explains this poor result

[15] This is based on the casual observation of age of consent, medical requirements, and waiting periods by state. Note that the two non-Catholic states with strict divorce laws—North and South Carolina—seem to have very liberal marriage laws

[16] I wish to acknowledge the contribution of Donald Bogue in the preparation of these data.

TABLE 7
REGRESSION OF FEMMAR, 1960 STATE DATA, EFFICIENT ESTIMATES

INDEPENDENT VARIABLE	AGE					AGE				
	15–19	20–24	25–29	30–34	35–39	15–19	20–24	25–29	30–34	35–39
RELWGCN .	−0.65 (2.5) [0.3]	4.2 (3.9) [1.1]	6.0 (3.0) [2.0]	4.0 (1.9) [2.1]	4.5 (1.7) [2.7]					
RELWGSS						−3.0 (2.0) [1.5]	4.2 (3.1) [1.4]	6.7 (2.3) [3.0]	4.3 (1.6) [2.7]	4.5 (1.3) [3.3]
SXRT	25.4 (6.5) [3.9]	13.2 (4.7) [2.8]	28.4 (7.2) [3.9]	42.7 (6.3) [6.8]	51.4 (6.0) [8.5]	21.3 (6.7) [3.2]	11.9 (4.7) [2.5]	23.7 (6.5) [3.7]	35.2 (6.2) [5.7]	42.1 (6.1) [6.9]
DCOST	0.10 (0.01) [7.7]	0.14 (0.02) [7.0]	0.04 (0.02) [2.4]	0.02 (0.01) [2.4]	0.01 (0.01) [1.7]	0.10 (0.01) [8.1]	0.14 (0.02) [7.2]	0.04 (0.02) [2.7]	0.03 (0.01) [2.7]	0.02 (0.01) [2.2]
HSEVAL	−0.50 (0.11) [4.6]	−0.43 (0.17) [2.5]	−0.57 (0.14) [4.1]	−0.44 (0.09) [4.7]	−0.24 (0.08) [3.2]	−0.49 (0.11) [4.6]	−0.49 (0.16) [3.0]	−0.62 (0.13) [4.7]	−0.45 (0.09) [4.8]	−0.27 (0.08) [3.5]
CONSTANT	−7.7 (9.5) [0.8]	46.7 (9.8) [4.8]	51.2 (9.1) [5.6]	42.7 (7.0) [6.1]	31.6 (6.5) [4.8]	0.46 (9.2) [0.1]	48.1 (8.3) [5.8]	54.3 (7.2) [7.6]	48.9 (5.8) [8.4]	40.1 (5.6) [7.1]

NOTE.—Standard error in parentheses. Absolute value of the coefficient estimate divided by its standard error shown in brackets.

TABLE 8
REGRESSION OF FEMMAR (WOMEN, 15 YEARS AND OLDER), 1960 STATE AND COUNTY DATA

Independent Variable	State Data	County Data	Elasticities from County Data
RELWGAE		2.93 (7.37)	0.091
RELWGSS	4.08 (2.57)		
SXRT	55.89 (5.67)	11.22 (4.66)	0.165
DCOST	0.022 (2.35)	0.041 (6.56)	0.026
HSEVAL	−0.220 (1.86)	−0.315 (6.65)	−0.061
AGE	0.548 (1.99)	1.30 (12.38)	
AGRI	0.005 (0.09)	0.092 (5.25)	0.010
NONWT	0.028 (0.93)	−0.117 (9.69)	−0.020
CONSTANT	−28.23 (1.26)	−34.19 (5.28)	
R^2	.72	.52	
F	12.66	79.56	
(df)	(8,39)	(7,520)	

NOTE.—t-values in parentheses.

married in the county in each age group was identical to the national average (see table 1). Counties differ in many other ways which may influence marital behavior, but only two factors, the importance of agriculture (AGRI) and the relative size of the nonwhite population (NONWT), will be included. The regression results for the county data are in table 8. For comparison, a similar regression for age-aggregated state data is included. The results are consistent with the predictions of theory and those reported above. As expected, the parameter estimates are much more accurate. An interesting finding is that, *ceteris paribus*, nonwhites have a lower propensity to marry. Of course, everything is not being held constant, since what is being measured here is the effect of the lower male/female relative wage that is known to exist among blacks.[17]

---

[17] An analysis of SMSA data disaggregated by race is in progress. The county data were also used for a preliminary analysis of education. In the South, the education of women has by tradition been oriented toward household production. Then higher female education with the schooling of males held constant should increase the gains to marriage in the southern states. The following results were obtained by including median male and female schooling completed (MALSCH and FEMSCH) in the equation for FEMMAR. The absolute value of the t-statistic is in parentheses:

$$
\begin{array}{l}
\text{FEMMAR} = \underset{(1.11)}{-6.8} + \underset{(4.98)}{1.8}\,\text{RELWGAE} + \underset{(4.91)}{9.4}\,\text{SXRT} + \underset{(4.81)}{0.04}\,\text{DCOST} \\
\qquad \underset{(5.69)}{-0.62}\,\text{HSEVAL} + \underset{(5.37)}{0.09}\,\text{AGRI} - \underset{(8.67)}{0.11}\,\text{NONWT} + \underset{(7.73)}{0.74}\,\text{AGE} \\
\qquad \underset{(0.26)}{-0.13}\,\text{MALSCH} + \underset{(4.22)}{1.8}\,\text{FEMSCH}, \; R^2 = .72, \; F(9{,}238) = 66.7
\end{array}
$$

TABLE 9
REGRESSIONS OF FEMMAR AND BIRRATE,
1960 COUNTY DATA, 2SLS (WOMEN, 15 YEARS AND OLDER)

Explanatory Variable	Dependent Variable: FEMMAR	Dependent Variable: BIRRATE
FEMMAR		0.33 (0.15) [0.335]
BIRRATE	−0.009 (0.02) [0.009]	
RELWGAE	3.03 (0.45) [0.094]	
SXRT	11.5 (2.5) [0.167]	
DCOST	0.04 (0.006) [0.026]	
HSEVAL	−0.3 (0.04) [−0.062]	1.1 (0.2) [0.221]
AGE	1.33 (0.12)	1.03 (0.05)
AGRI	0.095 (0.02) [0.011]	0.063 (0.041) [0.007]
NONWT	−0.11 (0.01) [−0.020]	0.09 (0.04) [0.016]
FEMEARN		−7.0 (1.0) [−0.263]
FEMSCH		−1.15 (0.39) [−0.188]
CONSTANT	−35.89 (7.46)	−8.2 (9.5)
F	16.2	11.7
(df)	(8,2)	(7,3)
α	0.06	0.06

NOTE.—Standard error in parentheses. Elasticity at mean in brackets.

The county data offer an opportunity to investuate another proposition mentioned by Becker. Suppose that a major component of family utility is the consumption of own child services.[18] Then a direction of causation running from marriage to fertility may not be appropriate. It is the demand for children which is most relevant, and observed marital behavior is partially a derived demand. Table 9 lists two-stage least-squares estimates from a two-equation model treating births and marriage simultaneously.[19] Here FEMMAR is a function of the birth rate

---

[18] This terminology arose in recent economic research on fertility. A compendium of this work can be found in T. W. Schultz (1973c).

[19] The first attempt to deal with simultaneous family decisions empirically was by Nerlove and Schultz (1970).

(BIRRATE) and vice versa.[20] The other explanatory variables in the birth rate equation are those suggested by the new economic approach to fertility. Clearly, the proportion of females married and, therefore, factors influencing the marriage market, contributes significantly to the explanation of observed fertility On the other hand, there is a small elasticity of the proportion of females married with respect to the birth rate This can be explained in two ways. First, child services may not, in fact, constitute a dominant portion of couples' consumption bundles. Alternatively, this small elasticity may be a statistical anomaly caused by comparison of a change in a stock to a change of a flow. Regressing the rate of new marriages on the birth rate may prove more satisfactory These results do, however, suggest the fruitfulness of the simultaneous analysis of family decisions

The decision to marry is only one of a set of interrelated family choices that influence both the observed proportion of women married and the variables chosen to explain this proportion In a sense, the marriage equation is partially a reduced form from a simultaneous system in which all of the right-hand variables are the result of choices made in the past This possibility has already been introduced with regard to the relative wage effect, and it applies to the other variables as well The sex ratio, for instance, is determined by migration decisions If unmarried females decide to leave a region (as is true for a rural population), then a greater proportion of the females remaining are married, not simply because the sex ratio is higher, but also because unmarried women have left. Much more complete data are required, however, before the nature of these reverse causations can be investigated

V. Conclusions

Data from the 1960 census yield substantial empirical support for the economic interpretation of the marriage market Two factors, the ratio of the sexes and the cost of divorce, influence observed marital behavior as expected This is also true of the gains to specialization, but the results are somewhat tenuous for teenagers The evidence also suggests the tentative conclusion that long-term economic growth may result in more people desiring to remain single Finally, it is shown that the analysis of simultaneous family decisions is empirically feasible

---

[20] Means and standard deviation of these and all other variables used are given in table 10.

TABLE 10
Means of Variables Used in Regression Equations

Variable	State Data, Ages						SMSA Data, Ages					County Data
	15t	15–19	20–24	25–29	30–34	35–39	15–19	20–24	25–29	30–34	35–44	
FEMMAR	65.9 (1.9)	15.1 (4.0)	67.1 (5.1)	83.5 (3.3)	86.1 (2.6)	85.7 (2.3)	14.1 (4.6)	66.4 (6.4)	84.5 (4.4)	87.3 (2.3)	85.9 (1.7)	65.8 (4.5)
RELWGSS	1.81 (0.14)											
RELWGCN	1.69 (0.12)								1.69 (0.13)			
RELWGAE												2.04 (0.37)
SXRT	0.96 (0.03)	1.01 (0.04)	0.95 (0.03)	0.96 (0.03)	0.96 (0.03)	0.95 (0.03)	0.97 (0.13)	0.89 (0.15)	0.95 (0.05)	0.95 (0.04)	0.95 (0.03)	0.98 (0.07)
DCOST	42.2 (24.0)								41.9 (24.4)			43.9 (22.4)
CATH	19.4 (12.0)											
HSEVAL	11.9 (2.7)								14.3 (2.9)			12.7 (4.5)
RENTAL									75.2 (8.5)			
AGE (marriage)	62.3 (1.4)											65.9 (2.0)
AGE (birth rate)												63.4 (8.3)
AGRI	6.7 (5.2)											7.4 (11.5)
NONWT	11.0 (8.4)											12.0 (12.2)
BIRRATE												65.6 (12.4)
FEMEARN												2.3 (0.7)
FEMSCH												10.7 (1.2)

Note.—Standard deviation in parentheses

Comment

T. Dudley Wallace

North Carolina State University

Freiden has been reasonably faithful to the Becker (1973) marriage model in his attempt to estimate the reduced-form equation explaining variation in the proportion of females married His observational units are regional subaggregates in cross-section data Thus, the Freiden paper represents a commendable attempt to evaluate the usefulness of the model

The major problem arising in going to the data is in translating and measuring full income Full income is a precise concept in the theory but is made up of two main components, one of which is ambiguous in its effect upon marriage The two components of full income are (1) nonlabor income, and (2) full labor earnings if both male and female were completely occupied in the market Variation in the first component, nonlabor income, is unambiguous Marriage should vary directly with it. Variation in "full wages" affects both implicit costs of time and full income and therefore is ambiguous Thus, full income, in sum, is ambiguous, and Freiden chose a proxy for full income that enters his regressions as a single variable

In all his regressions, Freiden obtained a negative sign on his proxy for full income and in many cases the *t*-statistics indicate that the results are not solely due to a lack of richness in the data

Returning to the Becker theory, male and female wages and nonlabor income all enter the reduced-form relationship explaining marriage Variational effects of wage ratios and nonlabor earnings are clear, variation in the wage sum (full earnings) is not Nevertheless, all three enter and have separate effects Therefore, it seems that Freiden should have introduced an additional explanatory variable, namely, the sum of male-female wages. According to the theory, the relative-wage results should hold up in the presence of full earnings If Frieden's proxy for full income is appropriate, its net effect, holding full earnings constant, should be positive. Given the generality of the theory that guided the empirical work, one cannot anticipate the sign on "full earnings" as measured by the sum of the wages Moreover, the data may not be rich enough to yield significant results on relative wages, full earnings, and full income Thus, such results would be revealed by the data It may also be the case that less generality in the theory would further clarify the role of full earnings in its relationship to marriage For example, I investigated a Cobb-Douglas production function with unit scale for the joint household [1]

The main additional result gained by this specificity was the delineation of the effect of a change in the sum of male and female wages (full earnings), holding the wage ratio constant For a Cobb-Douglas household production function, *an increase in full earnings due to* equiproportionate changes in the male and female wages reduces the gain from the marriage whenever the ratio of full earnings to nonlabor income is less than the ratio of labor cost to market-goods cost Conversely, when the source of full income from labor relative to nonlabor income is larger than the relative share of labor-to-goods cost in production, an equal increase in both wage sources increases the attractiveness of marriage, given the Becker framework Whether this type of inference, or something similar, holds for more general production functions and whether data can be found to test more specific hypotheses about the role of full earnings are open to question

Not considering product allocation within the household contributes to a failure to sort out variables into cost-returns (supply-demand) boxes that have been found so useful in most work in economics The impression left in both the Freiden paper and the Becker paper (1973) is that one can say something about the incidence of marriage but, because of symmetry, little can be said about such matters as the demand for wives by males or the supply of wives If one took the simple and obvious expedient of assuming equal sharing of household production, the demand for wives could be represented by the optimal joint-product function less twice the single-male production, and the demand for husbands by women could be similarly represented by joint production minus twice the single-female production These are the returns On the cost side there are costs of search, measured in some sense by the relevant sex ratios, and costs of correcting mistakes, as reflected by ease of divorce Whether one could construct meaningful models that are identifiable in this way, given the sorts of data which are available, is moot If such constructions are possible, the link between model and data would be clarified As an example, in both the Freiden paper and in Becker (1973), the sex ratio and the divorce-cost variables are brought in by ad hoc argument rather than as an integral part of the theory This makes for a *ceteris paribus* problem in interpreting results The equations that Freiden fits with data are reduced forms of two sorts of considerations

---

[1] For good reason, Becker wanted to avoid increasing returns to scale Scale less than or equal to unity leaves the burden of the theory to complementarity in production As Becker has stated (1973), Cobb-Douglas (C-D) implies zero productivity in the singles cases Our sense of description may thus be offended by using the C-D function, but if we accept the notion that marriage is more attractive, the larger the difference in joint and individual household production, the general theory is not crucially violated by zero product in the households of single individuals

What follows are a few additional comments that may have marginal value (1) Certainly the existence of marriage predates progressive income tax and joint returns Nevertheless, levels of marriage incidence in modern society should depend on tax structure (2) Divorce costs probably vary directly with the number of immature children; thus, the index used by Freiden is not so appropriate for his higher-age groups. Divorce costs in some states vary greatly, depending upon real assets Also, lowering the probability of dissolution of long-standing marriages is the effect of the household equivalent of on-the-job training specific to the individual household One might expect considerable differences in the analysis of survey data where one could obtain, for example, specific information about household wealth, number of dependent children, or length of time of current marriage Survey data should also be amenable to resolution of the importance of household productivity of the less intensive wage earner (3) The notion that the "likelihood" of marriage increases the greater the difference in joint and individual household production is appealing but somewhat lacking in rigor The basic Becker theory is completely deterministic Therefore, in speaking of "likelihood," one leaps from the basic theory into an area still to be explored Whether beginning with a probabilistic framework is of practical value is not clear, but it should be recognized that there is here a methodological problem which is not usually present in microeconomic modeling.

Lest the critical tone of these comments be misleading, it is worth saying again that carrying economic analysis into unexplored terrain is exciting and of potentially great value. The Becker theory and the Freiden empirics are steps into a new area for economists. The steps are worthwhile and will not end here

Benefits of Women's Education within Marriage

Lee Benham

University of Chicago

Studies of the returns to education have generally investigated the relationship between individuals' investments in formal education and on-the-job training and their labor-market productivity It is well recognized, however, that other factors besides formal education and training contribute to a person's effective stock of human capital (and hence to productivity), these factors include early childhood environment, parents' behavior (see Dugan 1969), and associations with other individuals There is general agreement, for example, that a child's development is affected by the ability and performance of peers in school At a later stage, a significant part of an individual's college and graduate education appears to result from association with fellow students, the more able students contributing to the education of all After formal schooling is completed, close associates are likely to continue to affect an individual's further educational development and to influence the rate of depreciation of the individual's stock of knowledge

One of the more persuasive explanations of the observed strong positive relationship between formal education and labor-market productivity is that, in addition to providing specific skills, formal education improves the individual's ability to acquire and assimilate information, to perceive and understand changing conditions, and to respond effectively [1] From

I thank Alexandra Benham for comments and substantial assistance throughout this study, Gary Becker, Solomon Polachek, Robert Michael, and T W Schultz for helpful comments, and Elayne Howard for analyzing the data The investigation was partly supported by PHS grant HS00080 from the National Center for Health Services Research and Development

[1] In his study of migration, Aba Schwartz (1968) found results consistent with the hypothesis that one effect of education is to reduce the cost of obtaining information The responsiveness of individuals to lifetime-earnings differentials was found to increase monotonically with education Finis Welch (1970) distinguishes between the worker effect and the allocative effect of education He argues "Much of the 'leverage' associated with added schooling is drawn from the dynamic implications of changing technology " That is, education enables an individual to adapt more rapidly to changing conditions

TABLE 4

ESTIMATED EFFECT OF WIFE'S AGE AT MARRIAGE ON HUSBAND'S EARNINGS
(t-STATISTIC IN PARENTHESES)

Sample	Dependent Variable	Husband's Education	Wife's Education	Husband's Experience	Husband's Experience Squared	Wife's Age at Marriage	Constant	R^2	N
1/1,000 Census, 1960* ...	ln EARN	0.076 (21.2)	0.032 (7.7)	0.042 (15.9)	−0.00070 (−12.8)	−0.0061 (−3.1)	6.995 (117.0)	.207	4,625
SEO, 1967* ..	ln EARN	0.065 (24.7)	0.036 (11.6)	0.036 (18.6)	−0.00064 (−16.0)	−0.0034 (−3.2)	7.337 (180.7)	.202	8,055
SEO, 1967†	ln EARN	0.066 (23.4)	0.032 (9.6)	0.037 (18.9)	−0.00070 (−16.4)	−0.0032 (−2.2)	7.354 (166.4)	.207	7,052

---

* See table 1 for description.

† Same as table 1 description except that families were included only if husbands were married only once

TABLE 5

ESTIMATED PERCENTAGE DIFFERENTIALS IN HUSBAND'S EARNINGS ASSOCIATED WITH WIFE'S PREMARRIAGE AND POSTMARRIAGE EDUCATION

(*t*-STATISTICS IN PARENTHESES)

Dependent Variable	Husband's Education	Wife's Years of Premarriage Education	Wife's Years of Postmarriage Education	Husband's Experience	Husband's Experience Squared	Constant	R^2	N
1) ln EARN. .	0 070 (17 5)	0 041 (10 3)	0 043 (6 1)	0 052 (17 3)	−0 001 (14 1)	6 625	.222	4,780
2) ln EARN for subset of husbands whose wives have some postmarriage education	0 079 (8 9)	0 043 (3 6)	0 044 (3 1)	0 077 (12 8)	−0 001 (10 5)	6 289	273	982

NOTE.—Data are from a 20 percent subsample of the 1/1,000 sample of the 1960 U S Census, further restricted to include only males of ages 18–64, married once with spouse present, husbands of once-married women, non-Negro, nonfarm residents, not in school, having positive number of weeks worked, positive earnings, and positive total income in 1959 Wives' years of pre- and postmarriage education were determined by comparing their age at marriage with an age assigned to their corresponding level of education completed Values were assigned as follows for years of wife's education ≤16 5, age at completion = years of education +6 For years of wife's education >16 5, age at completion = 22 5 + 2 (years of education − 16 5) (The highest category for schooling completed is 18 years) In the subsample used for (1) above, for husband's education wife's premarriage education, and wife's postmarriage education, means are 11 328, 11 056, and 0 341 and standard deviations are 3 521, 3 126, and 1 63, respectively In the subsample used for (2) above, the corresponding means are 12 589, 11 507, and 1 660 and standard deviations are 3 232, 3 930, and 3 274 education is at least as strongly associated with husbands' earnings as is their premarriage education[17]

Another variant of the selective-mating hypothesis states that a wife's education is only a proxy for other background characteristics of both spouses, such as intelligence and social class Should these latter variables be associated both with husband's earnings and with wife's own educational attainment, then the association noted between the wife's education and the husband's earnings may be spurious This argument has its counterpart in the recurring question of the extent to which an individual's education itself results in higher earnings versus the extent tc which education is merely a proxy for intelligence, social background and so forth

Data on background variables were not available in the samples use
for tables 1–5 However, this issue has been examined in another contex
by Duncan, Featherman, and Duncan (1972, p 178) in a study of th
relationship between the occupational status of husbands and the back
ground characteristics (including education) of both spouses[18] A multipl
regression was computed of husband's occupational status in 1956 on hi
background characteristics, including his education, his occupation
status at marriage and his father's occupational status, and his wife's bac
ground characteristics, including her education, her father's occupation
status, her intelligence, and a psychological index of her drive to g
ahead All the coefficients for wife's characteristics *except education* we
less than two times their standard error This is consistent with the hyp
thesis that the wife's education has a significant independent effect on t
husband's career development.

These results, however, are not directly comparable to those present
in tables 1–5, since occupational status rather than earnings is t
dependent variable By using another data set developed by Blau a
Duncan (1967), it is possible to obtain more directly comparable resu
by estimating the relationship between husband's income in 1961 a
both spouses' education, given several of the husband's backgrou

---

[17] An alternative explanation concerning the strong relationship between w
postmarriage education and husband's earnings is that husbands with higher earn
purchase more postmarriage education for their wives To examine this questio
sample was drawn of husbands of ages 20–30 having 12 or more years of school
were employed full time during the week of the 1960 U S Census survey For this gr
median income was calculated for husbands whose wives were in school and for t
whose wives were not in school Only 2 4 percent of the men in this sample had a wi
school, so results should be viewed with appropriate caution, but in all men's educa
categories except 12 years, the median income was *lower* for those men with wiv
school

[18] The data used for these estimates were for couples who had recently had a se
child at the time of the survey in 1957 (see Duncan et al. [1972, p. 34] for a more com
description).

characteristics, including occupational status of his first job and his father's occupation and education.[19]

The following three specifications were used to estimate the impact of including husband's background characteristics using these data:

$$\text{INCOME} = \alpha_1 + \beta_1 \text{ EDH} + \beta_2 \text{ EDW} + \mu,$$

$$\begin{aligned}\text{INCOME} = \alpha_1 &+ \beta_1 \text{ EDH} + \beta_2 \text{ EDW}\\ &+ \beta_3 \text{ EDHF} + \beta_4 \text{ OCHF} + \mu, \text{ and}\end{aligned}$$

$$\begin{aligned}\text{INCOME} = \alpha_1 &+ \beta_1 \text{ EDH} + \beta_2 \text{ EDW}\\ &+ \beta_3 \text{ EDHF} + \beta_4 \text{ OCHF} + \beta_5 \text{ OCHJ} + \mu,\end{aligned}$$

where INCOME = husband's income in 1961, EDH = husband's education, EDW = wife's education, EDHF = husband's father's education, OCHF = husband's father's occupation, and OCHJ = husband's occupational status on his first job. Due to the nature of the available data,[20] these specifications differ from the earlier ones in this paper in that income rather than ln earnings is the dependent variable and husband's work experience is not included.[21] Nevertheless, they provide qualitative estimates of the relative importance of husband's background characteristics.

Table 6 shows the relative impact of wife's education as compared with husband's education on husband's income (β_2/β_1) by age of wife. These estimates are consistent with those presented earlier. There is a substantial relationship between wife's education and husband's income in most years. For wives of ages 22–26, husband's income increases 37 percent as much with an increase in wife's education as with a similar increase in husband's own education. This ratio increases to 80 percent for wives ages 32–36 and then declines at older ages, a pattern similar to that shown in table 3.

When husband's father's occupation and education are included, the pattern is very similar.[22] Except for the oldest group, these results are not consistent with the hypothesis that the association between wife's education

---

[19] Data were collected from 20,700 respondents representing men in the U.S. civilian noninstitutional population between the ages of 20 and 64 in March 1962 (see Duncan et al. 1972, p. 32).

[20] Professor Otis Dudley Duncan generously provided the correlation matrices which he had developed for table 8.16 in Duncan et al. (1972, p. 240). The estimates shown here were calculated from those matrices.

[21] The consequences of excluding the husband's work-experience variables are less severe here since separate estimates are made for each of eight age-of-wife categories.

[22] Inclusion of father's education and occupation reduces the education coefficient for both spouses. The ratio of the husband's education coefficient when his father's education and occupation are included to the husband's education coefficient when they are excluded, by wife's age, is as follows: wife's age 22–26 (.91), 27–31 (.81), 32–36 (.80), 37–41 (.82), 42–46 (.87), 47–51 (.85), 52–56 (.85), and 57–61 (.80).

Comment

Finis Welch

City University of New York and National Bureau of Economic Research

I have been tempted to write a note on the social cost of the "marriage go-round ” In a nutshell, the idea is that more productive marriag partners are preferred and that, on balance, net productivity in marriag is positively related to education Increased education serves to increas an individual's chance of a good marriage, of marriage to a "more educated spouse If an individual's social product is independent of h spouse's education, then the marriage-go-round results in overinvestme in education Benham's view that a person's own productivity depends c attributes of the spouse is a clear alternative

My skepticism concerning Benham's hypothesis that a husbanc earning capacity is positively related to his wife's schooling is n qualitative but quantitative How important are these (very lik positive) cross-productivity effects? I was not surprised to hear of t positive association between wife's schooling and husband's earnings was surprised that most emphasis is on productivity instead of select mating

Benham points out that, since his measures are of annual earnings, th may be labor-supply effects submerged in the estimated relations between husband's earnings and wife's schooling He notes (correc that the direction of this effect is unclear on a priori grounds, but fi empirically that an extra year of wife's schooling has only a slig smaller percentage effect on husband's weekly wage than on ann earnings In tabulations using a different body of data, I have veri this result I will briefly describe these computations

I used the NBER-Thorndike sample of 4,519 white married males (present) who entered the U S Air Force in 1943 and who were exam as navigators and pilots at that time Earnings are for 1969. Experien the number of years since the individual was last in school The geom mean of earnings is over $13,000,[1] and logarithmic variance is 0 24.

---

[1] Earnings for 1969 are deflated to the 1958 value so as to allow life-cycle compa In the regressions below, I did not rescale this variable

The advantage of this sample is that data are available to permit explicit consideration of the possible effects of selective mating For example, information on husbands includes (1) schooling, (2) a measure of IQ, (3) father's schooling and occupation, and (4) religion It is clear that the sample is rarefied, and it is consequently hard to judge how much confidence should be placed in results contained within For example, essentially all respondents are white, completed high school, and on armed forces intelligence tests scored in the upper one-half of the distribution On the other hand, in one important dimension, this sample is superior to the national probability samples for isolating "pure" productivity effects of wife's schooling. It is likely that the World War II marriage market functioned somewhat differently from that in normal times—possibly a careful matching of family background was less important during the war If so, then in this sample, containing a disproportionately large number of men who were married during the war, there should be less "pollution" with covariance between wife's schooling and husband's family background [2] For example, Benham, in an unpublished table, shows that, in the U S Census sample, the zero-order correlation between husband's and wife's schooling is 0 65 In the NBER-Thorndike sample the correlation is 0 38

It is true that this lower correlation in the NBER-Thorndike sample could be the result of truncation in the husband's schooling variable, but it is important that if truncation reduces correlations it does not change biases due to omitted variables Yet, if the reduced correlation follows from reduced covariance between wife's schooling and omitted variables such as husband's IQ or family background, which may themselves affect earnings, then the reduced correlation reduces coefficient bias Table 1 reports regression results for this sample

Regression 1 is equivalent to regressions Benham presents in table 1A, and the results agree that there is a "significant" association between wife's schooling and husband's earnings There is a quantitative difference, however, while a year of schooling for the husband has the same estimated effect (6–7 percent) on earnings, the coefficient for wife's schooling is less than one-half as high in the NBER-Thorndike sample as in the samples used by Benham Regression 2 includes the length-of-marriage variables à la Benham's table 3 Here no evidence of a statistically significant relationship exists The computed F-statistic (2, 4,512 degrees of freedom) is only 1.16 In any case, it is hard to interpret Benham's results vis-à-vis length of marriage, for the marriage can itself be considered as a short "period of investment "

Regression 3 introduces a decomposition of the wife's schooling variable into two parts The first part includes those components of a wife's

---

[2] Roughly 17.5 percent of the NBER-Thorndike sample of men who had wives in 1969 had married those particular spouses between 1941 and 1944 Of this group of men, 8 percent were married in 1944.

TABLE 1
REGRESSION ESTIMATES OF RELATIONSHIP BETWEEN HUSBAND'S EARNINGS AND WIFE'S SCHOOLING
(*t*-RATIOS IN PARENTHESES)

A DEPENDENT VARIABLE LOG$_e$ (HUSBAND'S ANNUAL EARNINGS)

Independent Variables	Regression No 1	Regression No 2	Regression No 3
Intercept	7.04	7.02	6.45
Husband's schooling	0.067 (18.4)	0.067 (18.4)	0.061 (16.5)
Experience	0.014 (4.6)	0.015 (4.6)	0.013 (4.1)
(Experience)^2	−0.0002 (−2.7)	−0.0002 (−2.8)	−0.0002 (−2.4)
Wife's education			
a) Observed	0.015 (4.2)	0.015 (4.3)	
b) Predicted by husband's IQ and background			0.144 (11.9)
c) Observed-predicted			0.007 (1.9)
Length of marriage		−0.002 (−0.4)	
(Length of marriage)^2		0.00012 (0.8)	
R^2	113	114	137

B DEPENDENT VARIABLE WIFE'S SCHOOLING

Independent Variables	Regression No 4
Intercept	12.1
Husband's IQ	0.11 (6.4)
Husband's father's schooling	0.12 (13.5)
Husband's father's occupation status	
a) High	0.23 (3.3)
b) Low	−0.03 (—0.33)
Husband's religion	
a) Jewish	0.44 (3.2)
b) Catholic	−0.27 (−3.7)
R^2	078

schooling that are correlated with (and presumably would be predicted by) her husband's IQ, religion, and his father's schooling and occupation. The intention is simply that these variables would be among the most obvious predictors of wife's schooling, based on a sociological "network" or an assortive-mating frame of reference The second part of the wife's education is that which is orthogonal to the above list and is simply the residuals between observed schooling and schooling predicted by regression 4 It is this part of a wife's schooling that is independent of her husband's background, religion, and measured IQ which is the obvious candidate for the "pure" productivity effect posited by Benham But in these data there is little independent role for wife's education Most of the effect is channeled through the independent variables of regression 4

To summarize these results, note first that the effect of a year of a wife's schooling is 22 percent of the estimated effect of her husband's schooling Further, after "partialing out" effects of a husband's IQ and background on his wife's schooling, the net effect of the wife's schooling is reduced to slightly less than half its original level and the coefficient estimate is of marginal statistical significance. The estimated net effect of a wife's schooling is roughly one-tenth of the estimated effect of her husband's schooling and is only one-fifth of Benham's estimate

Does this then prove that the pure productivity effect is unimportant? Clearly not Aside from the peculiarities of this particular set of data, the identification problems surrounding the Benham hypothesis are probably insurmountable For example, consider models of selective mating If forces for selectivity exist in the absence of this cross-productivity (cross-fertilization?) effect, they are clearly magnified by its existence

Family Investments in Human Capital: Earnings of Women

Jacob Mincer

Columbia University and National Bureau of Economic Research

Solomon Polachek

University of North Carolina

I. Introduction

It has long been recognized that consumption behavior repre
mainly joint household or family decisions rather than separate dec
of family members Accordingly, the observational units in consum
surveys are "consumer units," that is, households in which inco
largely pooled and consumption largely shared.

More recent is the recognition that an individual's use of tim
particularly the allocation of time between market and nonr
activities, is also best understood within the context of the famil
matter of interdependence with needs, activities, and characteri
other family members More generally, the family is viewed
economic unit which shares consumption and allocates produc
home and in the market as well as the investments in physical and
capital of its members In this view, the behavior of the family unit
a division of labor within it. Broadly speaking, this division of l
"differentiation of roles" emerges because the attempts to promot
life are necessarily constrained by complementarity and sub
relations in the household production process and by com

Research here reported is part of a continuing study of the distribution
conducted by the National Bureau of Economic Research and funded by th
Science Foundation and the Office of Economic Opportunity This report has
gone the usual NBER review We are grateful to Otis Dudley Duncan, James
Melvin Reder, T W Schultz, and Robert Willis for useful comments, and
Borjas for skillful research assistance

advantages due to differential skills and earning powers with which family members are endowed

Though the levels and distribution of these endowments can be taken as given in the short run, this is not true in a more complete perspective Even if each individual's endowment were genetically determined, purposive marital selection would make its distribution in the family endogenous, along the lines suggested by Becker in this volume Of course, individual endowments are not merely genetic, they can be augmented by processes of investment in human capital and reduced by depreciation Indeed, a major function of the family as a social institution is the building of human capital of children—a lengthy "gestation" process made even longer by growing demands of technology

Optimal investment in human capital of any family member requires attention not only to the human and financial capacities in the family, but also to the prospective utilization of the capital which is being accumulated Expectations of future family and market activities of individuals are, therefore, important determinants of levels and forms of investment in human capital Thus, family investments and time allocation are linked while the current distribution of human capital influences the current allocation of time within the family, the prospective allocation of time influences current investments in human capital

That the differential allocation of time and of investments in human capital is generally sex linked and subject to technological and cultural changes is a matter of fact which is outside the scope of our analysis Given the sex linkage, we focus on the relation within the family between time allocation and investments in human capital which give rise to the observed market earnings of women Whether these earnings, or the investments underlying them, are also influenced or reinforced by discriminatory attitudes of employers and fellow workers toward women in the labor market is a question we do not explore directly, though we briefly analyze the male-female wage differential Our major purposes are to ascertain and to estimate the effects of human-capital accumulation on market earnings and wage rates of women, to infer the magnitudes and course of such investments over the life histories of women, and to interpret these histories in the context of past expectations and of current and prospective family life.

The data we study, the 1967 National Longitudinal Survey of Work Experience (NLS), afford a heretofore unavailable opportunity to relate family and work histories of women to their current market earning power Accumulation of human capital is a lifetime process In the post-school stage of the life cycle much of the continued accumulation of earning power takes place on the job. Where past work experience of men can be measured without much error in numbers of years elapsed since leaving school, such a measure of "potential work experience" is clearly inadequate for members of the labor force among whom the lengt and continuity of work experience varies a great deal. Direct informatio on work histories of women is, therefore, a basic requirement for th analysis of their earnings To our knowledge, the NLS is the only dat set which provides this information, albeit on a retrospective basi Eventually, the NLS panel surveys will provide the information on current basis, showing developments as they unfold.[1]

II. The Human-Capital Earnings Function

To the extent that earnings in the labor market are a function of human-capital stock accumulated by individuals, a sequence of posit net investments gives rise to growing earning power over the life cy When net investment is negative, that is, when market skills are erode depreciation, earning power declines This relation between the seque of capital accumulation and the resulting growth in earnings has formalized in the "human-capital earnings function " A simple spe cation of this function fits the life cycle "earnings profile" of men ra well. The approach to distribution of earnings among male wo (in the United States and elsewhere) as a distribution of indivi earnings profiles appears to be promising.[2]

For the purpose of this paper, a brief development of the ear function may suffice

Let C_{t-1} be the dollar amount of net investment in period $t - 1$, (gross) earnings in that period, before the investment expenditur subtracted, are E_{t-1} Let r be the average rate of return to the indivi human-capital investment, and assume that r is the same in each Then

$$E_t = E_{t-1} + rC_{t-1}$$

Let $k_t = C_t/E_t$, the ratio of investment expenditures to gross ea which may be viewed as investment in time-equivalent units Th

$$E_t = E_{t-1}(1 + rk_{t-1}).$$

---

[1] For a description of the NLS survey of women's work histories, see Par Spitz, and Zeller (1970) For an analysis of earnings of men, using "potent experience measures, see Mincer (1974) Though less appropriate, the same pro was used in several recent studies of female earnings Direct information fro Survey was first used by Suter and Miller (1971) The human-capital approa applied to these data by Polachek in his Columbia Ph.D thesis, "Work Exp the Difference between Male and Female Wages" (1973) This paper repo development of the analysis in that thesis

[2] See, for instance, Rahm (1971), Chiswick and Mincer (1972), Chisw Mincer (1974), and a series of unpublished research papers by George E J Frank P. Stafford on earnings of Ph D 's in various fields

By recursion $E_t = E_0(1 + rk_0)(1 + rk_1) \quad (1 + rk_{t-1})$ The term rk is a small fraction Hence a logarithmic approximation of $\ln\,(1 + rk) \simeq rk$ yields

$$\ln E_t = \ln E_0 + r \sum_{i=0}^{t-1} k_i \tag{3}$$

Since earnings net of investment expenditures, $Y_t = E_t(1 - k_t)$, we have also

$$\ln Y_t = \ln E_0 + r \sum_{i=0}^{t-1} k_i + \ln\,(1 - k_t) \tag{4}$$

Some investments are in the form of schooling, others take the form of formal and informal job training If only these two categories of investment are analyzed, that is, schooling and postschool experience,[3] the k terms can be separated, and

$$\ln E_t = \ln E_0 + r \sum_{i=0}^{s-1} k_i + r \sum_{j=s}^{t-1} k_j \tag{5}$$

where the k_i are investment ratios during the schooling period and the k_j thereafter With tuition added to opportunity costs and student earnings and scholarships subtracted from them, the rough assumption $k_i = 1$ may be used[4] Hence,

$$\ln E_t = \ln E_0 + rs + r \sum_{j=s}^{t-1} k_j \tag{6}$$

The postschool investment ratios k_j are expected to decline continuously if work experience is expected to be continuous and the purpose of investment is acquisition and maintenance of market earning power This conclusion emerges from models of optimal distribution of investment expenditures C_t over the life cycle (see Becker 1967 and Ben-Porath 1967) A sufficient rationale for our purposes is that as t increases, the remaining working life $(T - t)$ shortens Since $(T - t)$ is the length of the payoff period on investments in t, the incentives to invest and the magnitudes of investment decline over the (continuous) working life This is true for C_t and a fortiori for k_t, since with positive C_t, E_t rises, and k_t is the ratio of C_t to E_t

In analyses of male earnings, a linearly (or geometrically) declining approximation of the working-life profile of investment ratios k_t appears to be a satisfactory statistical hypothesis

---

[3] The inclusion of other categories in the earnings function is an important research eed, since human capital is acquired in many other ways in the home environment, in nvestments in health, by mobility, information, and so forth

[4] According to T. W Schultz, this assumption overstates k, especially at higher educaon levels, leading to an understatement of r

It will be useful for our purpose of studying earnings of women t decompose net investments explicitly into gross investments and d preciation Let C^*_{t-1} be the dollar amount of gross investment in peric $t-1$, δ_{t-1} the depreciation rate of the stock of human capital, hence earnings E_{t-1} during that period, and $k^*_t = C^*_t/E_t$, the gross investme ratio. Hence

$$E_t = E_{t-1} + rC^*_{t-1} - \delta_{t-1}E_{t-1}$$

and

$$\frac{E_t}{E_{t-1}} = 1 + rk^*_{t-1} - \delta_{t-1} = 1 + rk_{t-1}, \quad \text{by equation (2),} \quad ($$

thus

$$rk_t = rk^*_t - \delta_t.$$

The earnings function (3) can, therefore, be written as

$$\ln E_t = \ln E_0 + \sum_{i=0}^{t-1} (rk^*_i - \delta_i)$$

In transferring the analysis to women, we face two basic facts: (1) marriage, women spend less than half of their lifetime in the market, on average Of course, this "lifetime participation rate" by marital status, number of children, and other circumstances, has been growing secularly. (2) The lesser market work of married is not only a matter of fewer years during a lifetime, and fewer we year, or a shorter work week An important aspect is discontir work experience, for most of the married women surveyed reported several entries into and exits from the labor force after school.

The implications of these facts for the volume and the life-c tribution of human-capital investments can be stated briefly [5]

1 Since job-related investment in human capital commands which is received at work,[6] the shorter the expected and actual of work experience, the weaker the incentives to augment job the life cycle With labor-force attachment of married women average, about one-half that of men, labor-market activities are less likely to contain skill training and learning compo result both of women's own decisions and decisions of empl may be expected to invest in worker skills to some extent.

2. Given discontinuity of work experience, the conclusion ation analysis to the effect that human-capital investme

---

[5] For a mathematical statement of the optimization analysis applied to work experience, see Polachek (1973, chap. 3)

[6] For the sake of brevity, the term "work" refers to work in the job marl imply that women occupied in the household do not work.

TABLE 1
Labor-Force Participation of Mothers Proportion Working, White Married Women with Children, Spouse Present

Age	Proportion Working (%) In 1966	After First Child	Ever	Sample Size
30–34	43	64	82	925
$S < 12$	46	71	75	294
$S = 12$	43	63	84	446
$S > 12$	40	59	88	185
35–39	47	67	87	945
$S < 12$	45	66	82	336
$S = 12$	49	68	88	422
$S > 12$	47	67	92	187
40–44	53	70	88	1,078
$S < 12$	52	72	78	465
$S = 12$	54	70	91	446
$S > 12$	51	68	93	167

Source —NLS, 1967 survey
Note —S = years of schooling

continuously over the successive years of life after leaving school is no longer valid Even a continuous decline over the years spent in the job market cannot be hypothesized if several intervals of work experience rather than one stretch represent the norm

3 The more continuous the participation, the larger the investments on initial job experience relative to those in later jobs

Women without children and without husbands may be expected to engage in continuous job experience But labor-force participation of married women, especially of mothers, varies over the life cycle, depending on the demands on their time in the household as well as on their skills and preferences relative to those of other family members The average pattern of labor-force experience is apparent in tables 1–3, which are based on the NLS data reported by women who were 30–44 years of age at the time of the survey According to the data

1 Though less than 50 percent of the mothers worked in 1966, close to 90 percent worked sometime after they left school, and two-thirds returned to the labor market after the birth of the first child (table 1) Lifetime labor-force participation of women without children or without husbands is, of course, greater

2 Never-married women spent 90 percent of their years after they left school in the labor market, while married women with children spent less than 50 percent of their time in it In each age group, childless women, those with children but without husbands (widowed, divorced, or separated), and those who married more than once spent less time in the market than never-married women, but more than mothers married once, spouse present (table 2).

TABLE 2

WORK HISTORIES OF WOMEN AGED 30–44 BY MARITAL STATUS (AVERAGE NUMBER OF YEARS)

GROUP	VARIABLE										SAMPLE SIZE
	h_1	e_1	h_2	e_2	h_3	e_3	Σe	Σh	S	N_c	
White, with children											
Married once, spouse present	0.57	3.55	6.71	1.14	1.22	1.69	6.4	10.4	11.8	3.16	2,398
Remarried, spouse present	0.54	2.43	7.85	2.60	2.02	2.00	7.1	10.3	10.6	3.28	341
Widowed	1.11	4.25	9.37	1.51	1.44	2.56	8.4	11.9	12.0	2.44	45
Divorced	0.94	2.96	6.54	4.24	2.38	2.92	10.1	9.8	10.8	2.98	133
Separated	0.74	3.97	7.81	2.71	1.14	2.08	8.7	9.6	10.1	2.86	65
White, childless											
Married once, spouse present	1.01	5.18		4.39	3.35	4.90	14.5	3.3	11.7		147
Never married		7.08			1.46	7.48	14.5	1.5	12.9		153
Black, with children											
Married once, spouse present	1.12	3.00	7.12	2.95	2.14	3.26	9.1	10.3	10.0	4.59	563
Remarried, spouse present	0.96	2.44	7.43	4.93	2.05	3.36	10.7	11.7	9.6	4.22	170
Widowed and divorced	1.19	2.23	7.67	4.36	1.90	3.68	10.3	10.8	9.8	4.20	149
Separated	1.28	2.86	6.24	5.57	2.38	2.81	11.2	9.8	9.4	4.22	191
Black, childless											
Married once, spouse present	2.33	4.75		3.83	4.58	4.77	13.4	6.9	10.9		71
Never married		7.15			4.74	6.45	13.6	4.7	10.9		47

ot worked between school and first marriage, e_1 = years worked between school and first marriage (for never-marrieds, = years worked prior to current job), rth of first child, e_2 = years worked after h_2 prior to current job, h_3 = interval of nonparticipation just prior to current job, e_3 = years nonparticipation since school, S = years of schooling, N_c = no of children.

3. Table 3 shows the characteristic work histories of mothers,[7] spouse present (MSP), who represented over two-thirds of the women in the sample We show chronologically the length of nonparticipation (h_1) during the interval between leaving school and marriage, the years of market work between school and the birth of the first child (e_1), an uninterrupted period of nonparticipation, h_2, starting just before the first child was born, followed by e_2 and h_3, which sum intermittent participation and nonparticipation, respectively, and finally e_3, the present job tenure of women working at the time of the survey

It is clear from the tabulations that, after their schooling, the life cycle of married women features several stages which differ in the nature and degree of labor-market and home involvement There is usually continuous market work prior to the birth of the first child The second stage is a period of nonparticipation related to childbearing and child care, lasting between 5 and 10 years, followed by intermittent participation before the youngest child reaches school age The third stage is a more permanent return to the labor force for some, though it may remain intermittent for others In our data, which were obtained from women who were less than 45 years old, only the beginning of the third stage is visible

The following conjectures about investment behavior in each of these stages are plausible in view of the described patterns which are to some extent anticipated by the women

1 Prospective discontinuity may well influence many young women during their prematernal employment (e_1) to acquire less job training than men with comparable education, unless they do not expect to marry or have an overriding commitment to a work career

2 During the period of childbearing and child care, prolonged nonparticipation may cause the skills acquired at school and at work to depreciate Some revisions of expectations and of commitments may also take place [8] Little investment, if any, can be expected during the episodic employment period e_2

3 There is likely to be a stronger expectation of prospective continuity of employment after the children reach school age To the extent that the current job (e_3) is more likely to represent this more-permanent return to the labor force than e_2 does, strong incentives to resume investments in job-related skills should reappear

---

[7] The six intervals shown in table 3 are aggregated from eight available ones Both sets are described in the Appendix

[8] We are reminded by T W Schultz that erosion of market skills during periods of nonparticipation is likely to be associated with growth in nonmarket productivity If so, the longer the time spent out of the labor force the greater the excess of the reservation or "shadow" price over the market wage, hence the smaller the probability of subsequent labor-force participation

TABLE 2

WORK HISTORIES OF WOMEN AGED 30–44 BY MARITAL STATUS (AVERAGE NUMBER OF YEARS)

GROUP	VARIABLE										SAMPLE SIZE
	h_1	e_1	h_2	e_2	h_3	e_3	Σe	Σh	S	N_c	
White, with children											
Married once, spouse present	0.57	3.55	6.71	1.14	1.22	1.69	6.4	10.4	11.8	3.16	2,398
Remarried, spouse present	0.54	2.43	7.85	2.60	2.02	2.00	7.1	10.3	10.6	3.28	341
Widowed	1.11	4.25	9.37	1.51	1.44	2.56	8.4	11.9	12.0	2.44	45
Divorced	0.94	2.96	6.54	4.24	2.38	2.92	10.1	9.8	10.8	2.98	133
Separated	0.74	3.97	7.81	2.71	1.14	2.08	8.7	9.6	10.1	2.86	65
White, childless											
Married once, spouse present	1.01	5.18		4.39	3.35	4.90	14.5	3.3	11.7		147
Never married		7.08			1.46	7.48	14.5	1.5	12.9		153
Black, with children											
Married once, spouse present	1.12	3.00	7.12	2.95	2.14	3.26	9.1	10.3	10.0	4.59	563
Remarried, spouse present	0.96	2.44	7.43	4.93	2.05	3.36	10.7	11.7	9.6	4.22	170
Widowed and divorced	1.19	2.23	7.67	4.36	1.90	3.68	10.3	10.8	9.8	4.20	149
Separated	1.28	2.86	6.24	5.57	2.38	2.81	11.2	9.8	9.4	4.22	191
Black, childless											
Married once, spouse present	2.33	4.75		3.83	4.58	4.77	13.4	6.9	10.9		71
Never married		7.15			4.74	6.45	13.6	4.7	10.9		47

h_1 = years not worked between school and first marriage, e_1 = years worked between school and first marriage (for never-marrieds, = years worked prior to current job), following birth of first child, e_2 = years worked after h_2 prior to current job, h_3 = interval of nonparticipation just prior to current job, e_3 = years of nonparticipation since school, S = years of schooling, N_c = no of children

3 Table 3 shows the characteristic work histories of mothers,[7] spouse present (MSP), who represented over two-thirds of the women in the sample. We show chronologically the length of nonparticipation (h_1) during the interval between leaving school and marriage, the years of market work between school and the birth of the first child (e_1), an uninterrupted period of nonparticipation, h_2, starting just before the first child was born, followed by e_2 and h_3, which sum intermittent participation and nonparticipation, respectively, and finally e_3, the present job tenure of women working at the time of the survey

It is clear from the tabulations that, after their schooling, the life cycle of married women features several stages which differ in the nature and degree of labor-market and home involvement There is usually continuous market work prior to the birth of the first child The second stage is a period of nonparticipation related to childbearing and child care, lasting between 5 and 10 years, followed by intermittent participation before the youngest child reaches school age The third stage is a more permanent return to the labor force for some, though it may remain intermittent for others In our data, which were obtained from women who were less than 45 years old, only the beginning of the third stage is visible

The following conjectures about investment behavior in each of these stages are plausible in view of the described patterns which are to some extent anticipated by the women

1 Prospective discontinuity may well influence many young women during their prematernal employment (e_1) to acquire less job training than men with comparable education, unless they do not expect to marry or have an overriding commitment to a work career

2 During the period of childbearing and child care, prolonged nonparticipation may cause the skills acquired at school and at work to depreciate Some revisions of expectations and of commitments may also take place [8] Little investment, if any, can be expected during the episodic employment period e_2

3 There is likely to be a stronger expectation of prospective continuity of employment after the children reach school age To the extent that the current job (e_3) is more likely to represent this more-permanent return to the labor force than e_2 does, strong incentives to resume investments in job-related skills should reappear

---

[7] The six intervals shown in table 3 are aggregated from eight available ones Both sets are described in the Appendix

[8] We are reminded by T W Schultz that erosion of market skills during periods of nonparticipation is likely to be associated with growth in nonmarket productivity If so, the longer the time spent out of the labor force the greater the excess of the reservation or "shadow" price over the market wage, hence the smaller the probability of subsequent labor-force participation

TABLE 3

Work Histories of Married Women by Age, Education, and Current Work Status

Age and Category	h_1	e_1	h_2	e_2	h_3	e_3	Σe	Σh	N_e	Sample Size
30–34:										
Worked in 1966										
$S < 12$	1 93	2 37	5 80	3 18	2 20	1 90	7 45	9 93	3 42	135
$S = 12–15$	0 90	2 84	5 41	2 21	1 39	2 31	7 36	7.70	2 89	233
$S \geq 16$	0 37	2 57	2 65	2 22	1 22	2 00	6.79	4 24	2 39	35
Did not work in 1966, but worked since birth of first child										
$S < 12$	1 67	2 23	6 29	1 31	5 09		3 54	13 05	3 50	68
$S = 12–15$	0 81	2 90	4 65	1 23	4 75		4 13	10 21	3 49	93
$S \geq 16$	0 50	1 85	3 57	1 71	3 57		3 56	7 64	3 00	14
Has not worked since birth of first child										
$S < 12$	4 54	1 42	9 64				1 42	14 18	3 24	85
$S = 12–15$	2 28	3 21	7 93				3 21	10 21	3 03	211
$S \geq 16$	1 95	1 11	7 20				1 11	9 15	3 14	34
35–39										
Worked in 1966										
$S < 12$	1 94	2 78	7 98	3 47	2 78	3 40	9 65	12 70	3 37	152
$S = 12–15$	0 98	3 42	6 85	3 09	2 01	3 70	10 21	9 84	2 99	250
$S \geq 16$	1 01	2 95	4 72	2 04	1 25	5 46	10 45	6 98	2 72	43
Did not work in 1966, but worked since birth of first child										
$S < 12$	2 15	2 96	9 00	1 80	6 40		4 76	17 55	3 70	65
	1 20	3 74	7 42	1 18	5 94		4 92	14 56	3 51	101
		5 75	6 50	1 15	2 62		6 90	9 50	2 87	8
							3 54	17.76	3 58	113
								14 59	3 16	170
										26

TABLE 3 (*Continued*)

40–44										
Worked in 1966										
$S < 12$	2 41	3 29	10 38	3 94	2 95	4 93	12 16	15 74	3 18	240
$S = 12–15$	1 55	4 16	8 74	3 57	2 63	4 43	12 16	12 92	2.72	297
$S \geq 16$	0 93	3 20	6 89	3 06	1 86	4 89	11 15	9 68	3 65	29
Did not work in 1966, but worked since birth of first child										
$S < 12$	2 35	3 31	12 95	1 51	6 89		4 82	22 19	3 41	89
$S = 12–15$	1 39	3 68	10 43	1 24	8 23		4 92	20 05	3 36	82
$S \geq 16$	3 19	1 19	9 80	1 34	4 80		2 53	17 79	3 59	5
Has not worked since birth of first child										
$S < 12$	6 23	2 63	17 66				2 63	23 89	3 93	130
$S = 12–15$	3 36	4 88	15 12				4 88	18 48	3 12	141
$S \geq 16$	3 03	2 67	13 35				2 67	16 38	2 96	31

NOTE —See notes to table 2 for explanation of variables

These conjectures imply that the investment profile of married wom
is not monotonic There is a gap which is likely to show negative valı
(net depreciation) during the childbearing period and two peaks bef
and after The levels of these peaks are likely to be correlated for the sa
woman, and their comparative size is likely to depend on the degre
continuity of work experience The whole profile can be visualize
comparison with the investment profiles of men and of single won
For never-married women, stage 1 (e_1) extends over their whole wor
life, and the investment profile declines as it does for men To the ex
however, that expectation of marriage and of childbearing are str
at younger ages and diminish with age, investment of never-ma
women is likely to be initially lower than that of men. At the same
given lesser expectations of marriage on the part of the never-married
initial on-the-job investments exceed those of the women who ever
marry, while the profile of the latter shows two peaks

The implications for comparative-earnings profiles are clear (
investment ratios imply a steeper growth of earnings, while de
investment profiles imply concavity of earnings profiles. Hence, e
profiles of men are steepest and concave, those of childless womer
and those of mothers are double peaked with least overall growt

III. Women's Wage Equation

To adapt the earnings function to persons with intermitt
experience we break up the postschool investment term in eq
into successive segments of participation and nonparticipatio
occur chronologically In the general case with n segments we m
the investment ratio $k_i = a_i + b_i t$, $i = 1, 2, \ldots, n$, and

$$\ln E_t = \ln E_0 + rs + r\sum_{i=1}^{n}\int_{t_i}^{t_{i+1}} (a_i + b_i t)\, dt.$$

Here a_i is the initial investment ratio, b_i is the rate of change c
ment ratio during the ith segment: $(t_{i+1} - t_i) = e_i =$ dur
ith segment Note that in (7) the initial investment ratio
projected value at $t_1 = 0$, the start of working life In a wo
which occurs in later life there is likely to be less investme
earlier interval j, though more than would be observed if j
its gradient through the years covered by m. In this case, c
(7) will exceed a_j

Alternatively, a_j and a_m can be compared directly in th

$$\ln E_t = \ln E_0 + rs + r\sum_{i=1}^{n}\int_{0}^{e_i} (a_i + b_i t)$$

since a_i is the investment ratio at the beginning of the part

While the rate of change in investment b_i is likely to be negative in longer intervals, it may not be significant in shorter ones Since the segments we observe in the histories of women before age 45 are relatively short, a simplified scheme is to assume a constant rate of net investment throughout a given segment, though differing among segments The earnings function simplifies to

$$\ln E_t = \ln E_0 + rs + r \sum_i a_i e_i \tag{9}$$

Whereas $(ra_i) > 0$ denotes positive net investment (ratios), $(ra_i) < 0$ represents net depreciation rates, likely in periods of nonparticipation

The question whether the annual investment or depreciation rates vary with the length of the interval is ultimately an empirical one Even if each woman were to invest diminishing amounts over a segment of work experience, those women who stay longer in the labor market are likely to invest more per unit of time, so that a_i is likely to be a positive function of the length of the interval in the cross section

Thus, even if $k_{ij} = a_{ij} - b_{ij}t$ for a given woman j, if $a_{ij} = \alpha_j + \beta_j t$ across women, on substitution, the coefficient b of t may become negligible or even positive in the cross section On integrating, and using three segments of working life as an example, earnings functions (7), (8), and (9) become

$$\begin{aligned}\ln E_t = a_0 + rs + r[a_1 t_1 + \tfrac{1}{2}b_1 t_1^2 + a_2(t_2 - t_1) \\ + \tfrac{1}{2}b_2(t_2^2 - t_1^2) + a_3(t - t_2) + \tfrac{1}{2}b_3(t^2 - t_2^2)],\end{aligned} \tag{7a}$$

$$\begin{aligned}\ln E_t = a_0 + rs + r(a_1 e_1 + \tfrac{1}{2}b_1 e_1^2 + a_2 e_2 \\ + \tfrac{1}{2}b_2 e_2^2 + a_3 e_3 + \tfrac{1}{2}b_3 e_3^2),\end{aligned} \tag{8a}$$

$$\ln E_t = a_0 + rs + r(a_1 e_1 + a_2 e_2 + a_3 e_3) \tag{9a}$$

In this example, t is within the last (third) segment, and the middle segment, $e_2 = h$, is a period of nonparticipation or "home time " The signs of b_i are ambiguous in the cross section, as already indicated, the coefficients of e_1 and of e_3 are expected to be positive, but those of e_2 (or h) negative, most clearly in (9a)

The equations for observed earnings ($\ln Y_t$) differ from the equations shown above by a term $\ln(1 - k_t)$—as was shown in the comparison of equations (3) and (4) With k_t relatively small, only the intercept a_0 is affected, so the same form holds for $\ln Y_t$ as for $\ln E_t$

It will help our understanding of the estimates of depreciation rates to express earnings function (9a) in terms of gross-investment rates and depreciation rates

$$\begin{aligned}\ln E_t &= \ln E_0 + \sum_i (rk_i^* - \delta_i) \\ &= \ln E_0 + (rs - \delta_s) + (rk_1^* - \delta_1)e_1 \\ &\quad + (rk_h^* - \delta_h)h + (rk_3^* - \delta_3)e_3\end{aligned} \tag{9b}$$

This formulation suggests that depreciation of earning power may occur not only in periods of nonparticipation (h), but at other times as well. On the other hand, market-oriented investment, such as informal study and job search, may take place during home time, so that $k_h^* > 0$. Positiv coefficients of e_1 and e_3 would reflect positive net investment, while negative coefficient of h is an estimate of net depreciation. If $k_h^* > 0$, th absolute value of the depreciation rate δ_h is underestimated

IV. Empirical Findings

Tables 4–8 show results of regression analyses which apply our earn function to analyze wage rates of women who worked in 1966, the preceding the survey The general specification is ln $w = f(S, e, h, x)$ where w is the hourly wage rate; S is the years of schooling, e is a v of work-experience segments, h is a vector of home-time segments x is a vector of other variables, such as indexes of job training, mot health, number of children, and current weeks and hours of work the statistical residual

The findings described here are based on ordinary least-squares regressions The tables show shorter and longer lists of variables covering all the intermediate lists In view of a plausible simul problem we attempted also a two-stage least-squares (2SLS) est procedure, which we describe in the next section. Since the 2SLS e do not appear to contradict the findings based on OLS, we descri first below

1. *Work History Detail and Equation Form*

When life histories are segmented into five intervals (eight is the possible in the data), three of which are periods of work expe two of nonmarket activity,[9] both nonlinear formulations (equ [7] and [8]) are less informative than the linear specification of change in investment (coefficient b) are probably not substa a short interval, and the intercorrelation of the linear an terms hinders the estimation. Dropping the square terms explanatory power of the regression slightly but increases of the life-cycle investment profile Conversely, when the aggregated, the quadratic term becomes negative but d acquire statistical significance by conventional standards. T term for current work experience is negative and significa

---

[9] Tables 2 and 3 show six intervals, including a very short nonpartic between school and marriage This interval is aggregated in other regressions.

of never-married women, one segment of work experience usually covers most of the potential working life Here the nonlinear formulation over the interval is as natural and informative as it is for men

2 *Investment Rates*

Table 4 compares earnings functions of women by marital status and presence of children, tables 5 and 6 by level of schooling, and table 7 by lifetime work experience In each table we can compare groups of women with differential labor-force attachment According to human-capital theory, higher investment levels should be observed in groups with stronger labor-force attachment

We can infer these differences in investment by looking at the coefficients of experience segments, e_1 (prematernal), e_2 (intermittent, after the first child), and e_3 (current) These increase systematically from married women with children to married women without children to single women in table 4, and from women who worked less than half to those who worked more than half of their lifetime in table 7 An exception is the coefficient of e_3 which appears to be somewhat higher for the group who worked less (see table 7) Note, however, that these coefficients are investment ratios (to gross wage rates), not dollar volumes Since wage rates are higher in the groups with more work experience, the conclusions about increasing investment hold for dollar magnitudes, a fortiori, and the anomaly in table 7 disappears [10]

Classifications by schooling show mixed results In table 5, where schooling is stratified by <12, 12–15, and 16+, investment ratios (coefficients of e_i) are lower at higher levels of schooling (with the exception of the coefficient of e_1) Translated into dollar terms,[11] no clear pattern emerges At the same time in table 6, where the schooling strata are ≤ 8, 9–12, and 13+, a positive relation between investment volumes and levels of schooling is somewhat better indicated Note that the sample size for the highest-schooling groups (10+) is quite small in table 5, as is that for the lowest-schooling groups (≤ 8) in table 6

3. *Investment Profiles*

Another implication of the human-capital theory refers to the shape of the investment profile. it is monotonically declining in groups with continuous participation, hence earnings are parabolic in aggregated

---

[10] The coefficient of e_3, calculated as $\partial \ln W/\partial e$, is 15 percent higher in the right-hand group However, the wage rate of this group is about 25 percent lower

[11] Wage rates are roughly 30 percent higher in successive schooling groups

TABLE 4

EARNINGS FUNCTIONS, WHITE WOMEN

	WITH CHILDREN								NO CHILDREN		NEVER MARRIED	
Var	b (1)	t	Var	b (2)	t	Var	b (3)	t	b (4)	t	b (5)	t
C	38		C	21		C	09		− 42		55	
S	076	11 5	S	063	10 5	S	064	12 0	081	4 4	077	4.9
$(A-S-6)$	.014	3 8	e	012	1 6	e_1	008	2 8	014	1 6	026*	1.5
$(A-S-6)^2$	− 001	−4 2	e^2	− 0002	−0 5	e_2	001	0 3	011	1 3	− 0007†	−1 1
R^2	16		e_3	021	2 8	e_3	012	2 7	015	2 2	009	1 5
			e_3^2	− 0008	−1 9	h_1	− 012	−2 5	− 005‡	−1 5	− 009‡	−0 6
			h	− 007	−1 5	h_2	− 003	−0 7	002§	0 7		
			h^2	000	0 2	etr	0002	1 5	0003	2 4	0003	1 7
			R^2	25		ect	010	3 2	− 003	−1 2	− 011	−1 8
						hlt	− 0003	−1 3	− 002	−1 3	− 0008	−1.2
						res	001	1 2	006	1 7	− 012	−2.2
						loc	044	2 7	− 021	−0 4	− 02	−0 3
						ln Hr	− 11	−3 7	− 15	−1 6	− 43	−4 4
						ln Wk	03	1 6	25	2 2	21	1 4
						N_c	− 008	−1 0				
						R^2	28		39		41	
						N	993		147		138	

NOTE —Var = variable, C = intercept, S = years of schooling, A = age, e = total years of work, e_1 = years of work before first child, e_2 = years of work after first child, e_3 = current job tenure, h = total home time, h_1 = home time after first child, h_2 = other home time etr = experience × training (months), ect = experience × certificate (dummy), on of illness (months), res = years of residence in county, loc = size of place of residence at age 15, ln Hrs = (log of) hours of work per week on current job, ln Wks = (log of N_c = no of children, b = regression coefficient, t = t-ratio, R^2 = coefficient of determination, N = sample size

TABLE 5
EARNINGS FUNCTIONS OF WMSP, BY SCHOOLING

VAR.	$S < 12$				$S = 12\text{–}15$				$S = 16+$			
	b	*t*	*b*	*t*	*b*	*t*	*b*	*t*	*b*	*t*	*b*	*t*
C	– 095		– 98		– 61		– 03		86		.36	
S..	046	4 7	039	3 8	105	5 1	086	4 0	038	0 4	107	1 1
e_1	016	3 1	015	2 2	012	1 7	008	1 3	023	1 5	010	0 5
e_2	014	2 8	012	1 8	006	1 2	0	0	– 013	–2 3	– 016	–3.0
e_3	021	4 7	019	3 2	015	3 7	011	1 9	002	1 6	.004	2 4
h_1	– 002	–0 6	001	0 2	– 013	–3 4	– 018	–2 8	– 023	–2 2	– 012	–1 7
h_2	002	0 5	003	0 5	– 001	–0 2	– 006	–1 0	006	0 4	002	1 0
etr			0004	1 2			0004	2 0			0007	0 5
ect			016	2 4			008	1 9			032	2 1
hlt			– 0006	–1 7			0	0			0	0
res			001	0 5			002	1 3			008	1.2
loc			06	2 4			036	1 6			05	0 8
ln *Hrs*			– 044	–0 9			– 11	–3 9			– 16	–3 4
ln *Wks*			045	1 5			031	1 2			05	0 7
N_e			– 004	–0 4			– 002	–0 2			– 05	–1 5
R^2	17		22		14		18		16		39	
N	435				622				83			

NOTE —WMSP = white married women, spouse present See table 4 for key to symbols

TABLE 6
Earnings Functions of WMSP (with Children), by Schooling

Var	S ≤ 8				S = 9–12				S = 13+			
	b	*t*	*b*	*t*	*b*	*t*	*b*	*t*	*b*	*t*	*b*	*t*
S	049	1 6	044	1 3	051	3 2	055	3 4	.068	2.8	.079	2.7
e_1	007	0 4	– 002	–0 4	013	1 7	012	1 5	021	1 4	.018	1.2
e_2	–.004	–2 1	– 028	–1 8	009	1 6	003	0 6	– 020	–1.5	–.020	–1.4
e_3	–.002	–0 3	– 008	–0 5	013	0 7	009	0.5	.009	2 0	.011	2.2
h_1	– 011	–1 5	– 007	–1 2	– 014	–1 8	– 010	–1 6	–.043	–3 1	–.031	–2.8
h_2	– 006	–0 4	– 003	–0 2	– 002	–0 4	– 002	–0 4	– 005	–0.4	–.004	–0.3
hlt			– 0007	–0 7			– 0011	–2 3			–.009	–0.6
ln *Hrs*			– 050	–0 7			– 090	–1.8			–.130	–1.1
ln *Wks*			– 070	–0 6			060	1 6			.090	1.2
N_e			– 008	–0.2			– 019	–0 4			–.010	–2.0
R^2	26		32		21		.26		.27		.33	
N	182				593				218			

men spouse present See table 4 for key to symbols

TABLE 7
EARNINGS FUNCTIONS OF WMSP BY LIFETIME WORK EXPERIENCE

	WORKED MORE THAN HALF OF YEARS			WORKED LESS THAN HALF OF YEARS		
VAR.	*b*	*t*	*M*	*b*	*t*	*M*
	− 28			− 10		
.	073	9 4	11 8	059	7 9	11 0
.	009	2 1	4 9	003	0 4	2 2
	006	1 4	5 6	− 005	−0 6	1 5
	017	2 0	4 9	022	3 8	1 6
	− 0002	−0 7		− 001	−1 5	
	− 014	−2 3	2 2	− 010	−2 6	10 7
$_1$	011	1 7	2 1	− 004	−0 9	4 7
$_2$	− 0008	−2 1	10 8	− 0001	−0 3	13 7
hlt	002	1 1	12 1	002	1 0	11 8
res	064	2 8	0 97	024	1 0	0 90
loc	− 08	−2 0	3 52	− 13	−4 4	3 40
ln *Hrs*	07	1 9	3 71	023	1 0	3 29
ln *Wks*	− 015	−1 4	2 21	− 001	−0 2	3 18
N_c						
R^2	22			21		
N	536			604		

NOTE —WMSP = white married women, spouse present See table 4 for key to symbols

xperience for men and never-married women[12] In the groups with iscontinuous participation, the profiles are not expected to be monotonic We can summarize the implicit profiles schematically, in terms of the oefficients of e_1, length of work experience before the first child, h_1, ninterrupted nonparticipation after the first child, and e_3, the current ork interval We find (table 4, col 3) that white married women with ildren (with spouse present) have current investment (ratio which ceeds the investment (ratio) incurred in experience before the first ild[13] Presumably, current participation in the labor force, which es place when most of the children have reached school age, is expected last longer than the previous periods of work experience This is tainly true of women over age 35, and it holds in regressions with or hout standardization for age

ooking at regressions within three education levels (tables 5–6), we that coefficient of prematernal experience (e_1) exceeds the coefficient urrent work experience (e_3) at the highest level of schooling (in the t equations, though not in the long ones), and the opposite is true at r levels. For women without children the coefficient of prematernal experience equals that of current work experience The investment le of never-married women has a downward slope Comparable

---

n the earnings regressions, the quadratic term of aggregated experience is often e, but not significant statistically

l statements about differences in coefficients refer to point estimates The dif- are mentioned because they are suggestive, though they would not pass strict tatistical significance within a given equation

early segments of their post school job experience contain higher inves
ratios—indeed, the fit implies a linear decline of such ratios over t
cycle Evidently, women who intend to spend more time in the labc
invest more initially. This is true, presumably, even if their plans a
changed following marriage and childbearing.

4. *Depreciation Rates*

The coefficient of home time is negative, indicating a net depre
earning power During the home-time interval (h_1), associa
marriage or the birth of the first child, this net depreciation ar
on average, 1.5 percent per year. In table 5 the depreciation ra
(−0.2 percent) and insignificant for women with less than h
education, larger (−1.3 percent) for those with 12–15 years of
and largest (−2.3 percent) for those with 16+ years of sc
table 6, the net depreciation rate is −1.1 percent for wome
mentary schooling or less, −1.4 percent for women with some
and −4.3 percent for women with at least some colleg
differences probably account for the different estimates in th
The depreciation rate also appears higher in the group who
than half the years (table 7)

It would seem that the depreciation rate is higher when th
stock of human capital is larger. An exception appears in t
of women without children (married and single) with wo
dren The former have a lower depreciation rate Of cours
spend much less time out of market work, and some of thi
job-oriented (e.g, job search)

It is useful to return to the formulation (9b) of the e
for a closer analysis of the depreciation rates: $\ln E_t = \ln E$
$(rk_1^* - \delta_1)e_1 + (rk_h^* - \delta_h)h + (rk_3^* - \delta_3)e_3$. Our coe
time measures the depreciation rate only if market-ori
k_h^* is negligible This is likely to be true for the period o
period defined as h_1 in the regression (h_2 in the tabula

An interesting question is whether the depreciatio
nonparticipation is different from the depreciation tha
well The question is whether depreciation due to no
capital stock (atrophy?) exceeds the depreciation du
to aging(?). We are inclined to believe that deprecia
("getting rusty") is by far more important, particul
relatively young (below age 45). Moreover, the at
that depreciation due to nonparticipation is stron
oriented components of human capital acquired on
for the inborn, initial, or general components of th
If so, a fixed rate of "home-time depreciation" a

ccumulation of human capital would appear as a varying rate in the arnings function given the volume of other human capital, the larger the n-the-job accumulated component of human capital, the higher the bserved (applied to the total earning power) depreciation rate [14]

This may be an explanation of the observed higher depreciation rates higher schooling and experience levels of mothers In particular, there a positive relation between the coefficients of h_t (in absolute value) and e_t across schooling groups (table 6), experience groups (table 7), and ce groups (compare tables 4 and 8)

Effect of Family Size

family size and number of children currently present affect the cumulation of earning power beyond the effect on work experience? e answer is largely negative when numbers of children and some asures of their age are added to work histories in the equations, the ldren variables are negative but usually not significant statistically ir inclusion reduces the absolute values of the coefficients of experience of home time and does not add perceptibly to the explanatory power he regression Note, however, that the children variable does approach nificance in the relatively small groups of highly educated women les 5–6), and more generally among women with stronger labor-force chment (table 7) Possibly, shorter hours or lesser intensity of work to some extent, the preferred alternatives to job discontinuity

ormal Postschool Training

coefficients of experience, a_t, represent estimates of rk_t, where k_t is verage investment ratio across women over the segment and r is the ge rate of return Individual variation in k_t is not available to us ve some individual information, however, on months of formal job ng received after completion of schooling as well as on possession of ional certificates by, among others, registered nurses, teachers, and cians If the length of training and possession of a certificate are indexes of k, we may represent $a_t = a_0 + \beta \cdot tr$, where tr is the of training. The term $a \quad e$ in that equation becomes

$$(a_0 + \beta tr) \quad e = a_0 \cdot e + \beta(tr \quad e)$$

interaction term $(tr \cdot e)$ can be added to the equation, and if the is is correct, the coefficient β should be positive This is indeed

---

δ is the observed depreciation rate, δ_j the rate applicable to job-accumulated and H_0 the volume of other human capital, $\delta = (\delta_j H_j)/(H_j + H_0) =$ $/H_j)]$ With a fixed rate δ_j for all individuals, the larger H_j the larger δ

the case in most of our equations, confirming the training interp
of the experience coefficients in the earnings function. Both inte
with months of job training and with possession of a certifi
significant for married women The training interaction variab
positive in the earnings function of single women, but the c
variable is negative. Whereas the negative coefficient of the cer
experience variable implies less than average investment behav
persons who work continuously, the corresponding positive coe
intermittent workers implies more than average investment be

7 *Effects of Mobility*

Research in mobility has shown that, so long as mobilit
voluntary—resulting from layoffs—it is associated with a gain
However, geographic labor mobility of married women is ofte
due to job changes of the husband. In that case, it may m
continuity of experience and slow the accumulation of ea
We used the information on the length of current residen
or a Standard Metropolitan Statistical Area (SMSA)
measure of mobility This variable has a small positive
rates of white MSP women and a significant negative e
women To the extent that mobility is job oriented fo
and exogenous for married women, the differential signs
sistent interpretation.

8 *Hours and Weeks in Current Job*

When (logs of) weeks and hours worked in the survey
in the regression, a negative sign appears for the weekl
and a positive but less significant one for the weeks-
The hours' coefficients are smaller for married wom
women and smaller for white than for black women.
of weekly hours may be partly or wholly spurious sin
indicated by respondents were weeks or months and t
was obtained by division through hours Of cour
causality is suspect: it is more likely that women w
work longer hours than the converse. Deletion of th
has a minimal effect on the equations.

9. *Other Variables*

Three other variables were included in the equati

1. Twenty percent of the married women who
out of work in 1967. We used a dummy variabl

TABLE 8
EARNINGS FUNCTIONS OF BLACK WOMEN

MSP WITH CHILDREN Var	b	t	NEVER MARRIED Var	b	t
C	− 02		C	− 48	
S	095	11 2	S	110	3 7
$_1$	005	0 8	e	004	0 1
$_2$	001	0 3	e^2	− 0003	−0 2
$_3$	.006	1 4	e_3	001	0 2
$_1$	− 006	−1 2	h	− 02	− 05
$_2$	− 005	−0 9	h^2	001	1 1
tr	0005	1 3	etr	0006	1 4
ct	008	1 9	ect	003	0 4
lt	− 0002	−0 5	hlt	− 001	−1 8
s	002	0 9	res	001	0 2
c	11	4 0	loc	23	2 7
Hrs	− 30	−7 4	ln Hrs	− 13	−0 7
Wks	08	2 2	ln Wks	03	0 2
$_c$	005	0 6	N_o		
^{2}	39		R^2	46	
	550		N	70	

NOTE — MSP = white married women, spouse present See table 4 for key to symbols

orking in 1966 stopped working in 1967, and 0 otherwise[15] This
iable had a negative sign, since it indicated a shorter current job
erience compared with the prospective work interval of others who
ntinued to work in 1967—the completed interval of those dropping out
s not longer than the interval of stayers In effect, women who dropped
of the labor force in 1967 had wage rates about 5 percent lower than
men who continued working, given the same characteristics and
ories[16] The proportion of dropouts is somewhat larger at lower
cation levels

. The size of community in which the respondent lived at age 15 had
ositive effect on earning power of married women but no effect on
of single women

Duration of current health problem in months was used as a measure
alth levels It is an imperfect measure for retrospective purposes and
s a very small negative effect on the wage rate

Black Women

gressions for black MSP (table 8) show experience coefficients about
e size of the corresponding white population Home time or
ation coefficients are not significant, neither are the children

---

hown in the tables

ut standardization, women who had dropped out had wage rates about 10
ver than women who continued working.

variables The implication is that there is less investment on the j
though black women spent more time than white women in t
market. They had more and younger children, on average T
variables behave comparably with those in the white regressio
that hours of current work and location at age 15 show strong
In contrast to white women, the size of community of residence
has a positive effect for never-married women as well. Agai
perience coefficients are smaller for black single women than
Perhaps contrary to expectations, neither health problems r
withdrawal from the labor force in 1966 differ for black as c
white married women with children, spouse present Rates
schooling appear, if anything, to be higher for black women

V. Lifetime Participation and the Simultaneity Prol

The earnings function, as we estimate it, relates wages
investments in schooling and on-the-job training and to
additional variables already discussed

The interpretation of some of the independent varia
affecting earning power may be challenged on the ground
just as well be viewed as effects rather than causes of
Presumably, women with greater earning power ha
aspirations and work commitments than other women
lifetimes Hence, what we interpret as an earnings func
read with causality running in the opposite direction—
function This argument is most telling for concurrent
last year's hours and weeks worked in relation to last
But these variables are of only marginal importance in
of married women All other independent variables t
the dependent variable (current wage rate), which
function interpretation less vulnerable, though not en
a serial correlation between current and past work ex
and past earning power Since lifetime work experier
on prior wage levels and expectations, our experie
part, *determined* as well as *determining* If so, the residu
tions is correlated with the experience variables,
coefficients which we interpreted as investment rati

How serious this problem is for our analysis de
of individual correlations between current and
tations of earning power and on the strength of eff
on subsequent work histories of individuals. Of co
grouped these correlations and effects are likel
educated women tend to have higher wage r
women throughout their working lives, (see, fo
our table 3 shows, they spend a larger fra

labor force. Table 3 also shows that married mothers who currently do not work, spent, on average, less of their lifetime working than those who currently work

One econometric approach to an estimation of the earnings function in the presence of endogeneity of "independent" variables is the two-stage least-squares (2SLS) approach We estimate work experience as a variable dependent on exogenous variables, some of which are in the earnings function and others outside of it In effect, we estimate a "lifetime labor-supply function " The second step is to replace the work-experience variables (e) in the earnings function by the estimated work experience($\hat{e}$) from the labor-supply function Parameter estimates in this revised earnings function are theoretically superior to the original, simple least-squares estimates[17]

Our application of a 2SLS procedure is far from thorough, for two reasons

1 It is difficult to implement it on the segmented function, since each of the segments would have to be estimated by exogenous variables For this purpose we aggregate years of work experience and compare the reestimated earnings function with the original, using aggregated experience

2 One of the variables in our lifetime labor-supply function is the number of children, which is not exogenous In principle, we should xpand the equation system to three to include the earnings function, the bor-supply function, and the fertility function At this exploratory level e prefer not to do it, particularly since the fertility function would be timated by the same variables as the labor-supply function

The supply function obtained for all white MSP women was

$$\frac{e}{e_p} = \underset{}{514} + \underset{(5\ 1)}{.020\ S_P} - \underset{(1\ 8)}{0064\ S_M} - \underset{(12\ 0)}{062\ N_c},$$

ere e is total years of work, e_p is "potential job experience," that is, rs since school, S_P is education of wife, S_M is education of husband, N_c is number of children The addition of earnings of husband uced the coefficient of S_M to insignificance without changing the ficient of determination, which was R^2 = 14

stimated values of the numerator ($\hat{e}$) are used to reestimate the earn- function A comparison of 2SLS and OLS estimates of the earnings ion is shown in table 9 If anything, the reestimated function shows r positive coefficients for (total) experience and stronger negative cients for home time The children variable becomes even less cant (in terms of t-values) than before The reestimation leaves our sions, based on the OLS regressions, largely intact

---

ce $\hat{e}$ is a function of exogenous variables, it is not correlated with the stochastic the reestimated earnings function

TABLE 9
EARNINGS FUNCTION, WMSP WOMEN, OLS AND 2SLS

VAR	OLS		2SLS		OLS		2SLS	
	b	*t*	*b*	*t*	*b*	*t*	*b*	*t*
C	−.20		−.06		.19		.26	
S	.069	12.8	.063	12.0	.053	9.4	.048	8.5
e	.010	3.2	.012	2.7	.008	2.8	.010	1.9
h_1	−.008	−3.0	−.015	−7.7	−.007	−1.9	−.013	−5.5
h_2	.0006	0.2	−.006	−2.3	.001	0.5	−.006	−1.9
e_3	.009	3.2	.009	3.5	.009	3.4	.010	3.7
tr					.005	2.2	.006	2.2
cert					.18	5.1	.18	5.1
hlt					−.0003	−1.3	−.0003	−1.4
res					.001	1.3	.021	1.4
loc					.044	2.8	.042	2.5
ln *Hrs*					−.11	−5.0	−.11	−4.9
ln *Wks*					.03	1.5	.03	1.6
N_c					−.010	−1.3	.003	0.3

NOTE.—WMSP = white married women, spouse present, *tr* = months of training, *cert* = certification (dummy), see table 4 for key to other symbols.

VI. Prediction

A test of the predictive power of the earnings function was performed on a small sample of women who did not work in 1966 but were found in the same first NLS survey to have returned to work in 1967. They were not included in our analyses, but their life histories and 1967 wage rates are available. The latter were predicted with several variants of the earnings function and compared to the reported wage rates. On average, the prediction is quite close, and the mean-square error is even smaller—relative to the variance of the observed wage rates—than the residual variance in the regressions.[18] In other words, the predictive power outside the data utilized for the regressions is no smaller than within the regressions. The test, however, is weak, because the sample is so small (45 observations). Similar tests will be performed on larger samples of women who return to the labor market in subsequent surveys.

VII. Earnings Inequality and the Explanatory Power of Earnings Functions

As table 10 indicates, the earnings function is capable of explaining 25–30 percent of the relative (logarithmic) dispersion in wage rates of white married women and about 40 percent of the inequality in the rather small sample of wage rates of single women in the 30–44 age group who worked in 1966. The earnings function is thus no less useful in understanding the structure of women's wages than it is in the analysis of wages of males.

---

[18] The (squared) correlation between predicted and actual wage rates was .37. The mean of actual rates was 5.196, with $\sigma = .335$; the mean of predicted wages was 5.187, with $\sigma = .204$.

TABLE 10
EARNINGS INEQUALITY AND EXPLANATORY POWER OF WAGE FUNCTIONS, 1966

Group	σ^2 (ln W)	R^2_W	σ^2 (ln Y)	R^2_Y	σ^2 (ln H)	N
Married women by education (yrs)						
< 12	17	21	81	76	64	435
12–15	18	17	92	78	74	622
+ 16	.17	16	77	74	60	83
Total	22	28	97	78	75	1,140
Single women	30	41	62	66	32	138
Married men	32	30	43	50	11	3,230

NOTE —σ^2 (ln W) = variance of (log) wages, σ^2 (ln Y) = variance of (log) annual earnings, σ^2 (ln H) = variance of (log) annual hours of work, R^2_W = coefficient of determination in wage rate function, R^2_Y = coefficient of determination in annual earnings function

The dispersion of hours worked during the survey year is much greater among married women, σ^2 (ln H) = 75, than among men, σ^2 (ln H) = 11 The (relative) dispersion in annual earnings of women is, therefore, dominated by the dispersion of hours worked This factor is also important in the inequality of annual earnings of single women and of men of comparable ages, but much less so It is not surprising, therefore, that the inclusion of hours worked in the earnings function raises the coefficient of determination from 28 percent in the hourly-wage equation to 78 percent in the annual-earnings equation of married women, from 41 percent to 66 percent for that of single women, and from 32 to 50 percent for that of men

The lesser inequality in the wage-rate structure of working married women than in the structure of male wages is probably due to lesser average, and correspondingly lesser variation in, job investments among individuals At the same time, the huge variation in hours, reflecting intermittency and part-time work as forms of labor-supply adjustments, creates an annual earnings inequality among women which exceeds that of men However, the meaning of that inequality, both in a causal and in a welfare sense, must be seen in the family context As was shown elsewhere (Mincer 1974), the inclusion of female earnings as a component of family income narrows the relative inequality of family incomes compared with that of incomes of male family-earners

VIII. Some Applications

1 *The Wage Gap*

To compare wage rates of women with wage rates of men, we analyzed earnings of men from the Survey of Economic Opportunity (SEO) for the same year (1966). We find that the average wage rate of white married men, aged 30–44, was $3.18, compared with $2 09 for white married women and $2.73 for white single women in our NLS data

TABLE 11
EXPERIENCE AND DEPRECIATION COEFFICIENTS, 1966, AGES 30–44

VAR	MARRIED WOMEN		SINGLE WOMEN		MARRIED MEN	
	b	*M*	*b*	*M*	*b*	*M*
S	063	11 3	077	12 5	071	11 6
$\hat{e}$	012	9 6				
e			.026	15 6	.034	19 4
e^2			−.0006	258	− 0006	409
e_3	009	3 2	009	8 0		
h_a	− 015	6 7				
h_0	− 006	3 5				

SOURCES —Women NLS 1967, men SEO, 1967
NOTE —S = years of schooling, h_a = home time following birth of first child, h_0 = other home time, e = years of work experience since completion of schooling, e_3 = current job tenure, $\hat{e}$ = 2SLS estimate of total work experience, b = regression coefficient, M = means

TABLE 12
EFFECTS OF WORK EXPERIENCE ON WAGE RATES

	RELATIVE CONTRIBUTION OF		PERCENT OF WAGE GAP EXPLAINED	
	Actual Experience (1)	Men's Experience (2)	(3)	(4)
Married women	+ 02	+ 26	45	42
Single women	+ 32	+ 33	7	40
Married men	+ 42			

We inquired to what extent the larger wage ratio (152 percent) of married men to married women and the smaller one (116 percent) of married men to single women can be explained by differences in work histories and by differences in job investment and depreciation For thi purpose we estimated a single earnings function of men, aged 30–44, in SEO The coefficients and means of the variables for these men are shown in table 11, which also gives the NLS estimates for both married and single women

Note that married men and married working women have just abou the same average schooling, while never-married women are somewha better educated (by 1 year, on average) The coefficients of schooling ar somewhat lower for married women but higher for single women Th big differences are in years of work experience since completion o schooling These are 19 4 for men, 15.6 for single women, and 9 6 fo married women The coefficients of initial experience are .034 for men .026 for single women, and about half as much for married women.

Multiplying the coefficients by the variables (table 11) and summin yields contributions of postschool investments to the (log of) wage rat as shown in table 12 These differences, roughly 40 percent betwee husbands and wives and 10 percent between married men and sing women, are about 70 percent of the observed difference in wage rates between married men and married women and a half of the difference between married men and single women.

If one prefers to be agnostic about the human-capital approach, one can treat the earnings function simply as a statistical relation and the regression coefficients as average "effects" of work experience and of nonparticipation on wages, without reading magnitudes of investment or depreciation into them. In that case we may ask how much the sex differential in wage rates would narrow if work experience of women were as long as that of men, but the female coefficients remained as they are. A multiplication of the female coefficients by the male variables in table 11 yields the following answers: for married women, 45 percent of the gap would be erased, for single women, only 7 percent of the much smaller gap (table 12, col. 3). The answer is similar for married women if the converse procedure is used, that is when the work experience of women is multiplied by the male coefficients (table 12, col. 4). For single women, the reduction of the gap is larger than in the first procedure.

We believe, however, that the weight of the empirical analysis of female earnings supports the view that the association of lower coefficients with lesser work experience is not fortuitous: a smaller fraction of time and energy is devoted to job advancement (training, learning, getting ahead) per unit of time by persons whose work attachment is lower. Hence, the 45 percent figure in the explanation of the gap by duration-of-work experience alone may be viewed as an understatment.

Indeed, comparing the annual earnings of year-round working women and men in the 30–40 age groups, Suter and Miller found a female-to-male earnings ratio of 46.7 percent. However, the ratio rose to 74 percent for women in this group who worked all their adult lives. The same comparison for high school educated persons yielded 40.5 as against 74.9 percent. Thus lifelong work experience reduces the wage gap by 51 or 58 percent, respectively.[19]

At this stage of research we cannot conclude that the remaining (unexplained) part of the wage gap is attributable to discrimination, nor, for that matter, that the "explained" part is not affected by discrimination. More precisely, we should distinguish between the concepts of direct and indirect effects of discrimination. Direct market discrimination occurs when different rental prices (wage rates) are paid by employers for the same unit of human capital owned by different persons (groups). In this sense, the wage-gap residual is an upper limit of the direct effects of market discrimination. Indirect effects occur in that the existence of market discrimination discourages the degree of market orientation in the expected allocation of time and diminishes incentives to investment in market-oriented human capital Hence, the lesser job investments and greater depreciation of female market earning power may to some extent be affected by expectations of discrimination

---

[19] Suter and Miller (1971, table 1). Their figures are not quite comparable with ours: their male data come from the Current Population Survey (CPS), and ours from SEO. They compare full-time earnings rather than wage rates, and they compare men and women without regard to marital status.

Of course, if division of labor in the family is equated with discrimination, all of the gap is by definition a symptom of discrimination. Otherwise, the analyses of existing wage gaps and of their changes over time remain meaningful, not tautological

Our data on work histories show some interesting trends which suggest a prospective narrowing of the wage differential. Table 3 shows that the uninterrupted period of nonparticipation which starts just prior to the birth of the first child has been shrinking when older women are compared with younger ones Women aged 40–44 who had their first child in the late 1940s stayed out of the labor force about 5 years longer than women aged 30–34 whose first child was born in the late 1950s Family size is about the same for both groups, but higher for the middle group (35–39) whose fertility marked the peak of the baby boom Still, the home-time interval in that group is shorter (by about 2 years) than in the older group and longer than in the younger Thus, the trend in labor-force participation of young mothers was persistent If, by the time the 30–34-year-old women get to be 40–44 (i e , in 1977), they will have had 4 years of work experience more than the older cohort, and their wage rates will rise by 6 percent on account of lesser depreciation and by another 2–4 percent due to longer work experience Thus, the total observed wage gap between men and women aged 40–44 should narrow by about one-fifth, while the gap due to work experience should be reduced by one-quarter[20]

2 *The Price of Time and the Opportunity Costs of Children*

The loss or reduction of market earnings of mothers due to demands on their time in child rearing represents a measure of family investment in the human capital of their children This investment cost has been measured by valuing the reduction of market time at the observed wage rate As pointed out by Michael and Lazear (1971), this valuation is incomplete for two reasons. First, if job investments take place at work, the observed wage rate understates the true foregone wage (gross or capacity wage) by the amount usually invested during the period when earnings are foregone. Second, as is clear from earnings-function analysis, the reduction of market time in turn reduces future wage rates because of a depreciation in earning power during the period of nonparticipation. The present value of future earnings lost through depreciation is a component of the opportunity cost of time, hence of children.[21]

---

[20] Two opposing biases mar this conjecture The shorter home-time interval for younger women is an average duration for those who already returned to work. It will lengthen with the passage of time as additional women return to the labor force It can be shown, however, that the apparent trend is genuine At the same time, the assumption of unchanged job-investment behavior leads to an understatement.

The data and the estimated wage functions permit a tentative, perhaps only an illustrative, empirical assessment of the opportunity cost of women's time and of children. Specifically, the marginal opportunity cost per hour of a year spent at home—rather than in the market—consists of (1) the gross wage rate (W_g), that is the observed but foregone wage (W) augmented by currently foregone investment costs, and (2) the present value of the reduction of the future gross wage through current depreciation.[22]

1. We can estimate W_g since $W = W_g(1 - k)$, and rk is estimated in the earnings function by a_1, the coefficient of work experience (e_1) preceding the interruption $k = a_1/r$, where r is the rate of return.

2. The present value of the reduction in W_g due to depreciation, using r as the discount rate, is $d/r \cdot W_g$, where d is the (depreciation) coefficient of home time in our wage equations.[23]

The estimates of marginal opportunity costs of a year (in dollars per hour) are shown in panel I of table 13 for three education groups of white mothers, aged 35–39. In panel II we calculate total opportunity expenditures incurred during the nonparticipation period following the birth of the first child. This is the period for which the earnings functions show significant depreciation coefficients. The length of the period depends, in part, on the number of children. Though interpreting all of the foregone earnings this period as an opportunity expenditure on children may be an overstatement, we impose an opposite bias by ignoring subsequent periods of non participation[24] which may also be child induced. Figures in panel II are the marginal costs per hour (per year) multiplied by h, the duration of home time. Figures in panel III are average opportunity expenditures per child (N_c) in each group. Since h is in years, the dollar figures in panels II and III should be multiplied by annual hours of work. For example, with 1,500 hours of work per year, the opportunity investment expenditures per child range from about $8,000 spent in 8.8 years by mothers with less than high school education to $17,000 spent in 5.2 years by mothers with college education or more.

---

[21] As Robert Willis suggested to us, this is strictly correct for the excess of depreciation during home time over the depreciation at other times. As we stated earlier, we believe that the latter is negligible in our age groups.

[22] Note that we are looking at household productivity as the return, the purpose of reducing market work, not as a negative element in costs.

[23] A 10 percent discount rate was used in these calculations.

[24] Inclusion would lead to a 20-25 percent increase in expenditures for the age group.

TABLE 13

MARGINAL PRICE OF TIME AND OPPORTUNITY COSTS OF CHILDREN, 1966, WORKING WHITE MOTHERS (AGED 35–39)—BY YEARS OF SCHOOLING
($)

	PRICE OF TIME PER HOUR			TOTAL OPPORTUNITY EXPENDITURES ON CHILDREN			OPPORTUNITY EXPENDITURES PER CHILD		
	< 12	12–15 (I)	16+	< 12	12–15 (II)	16+	< 12	12–15 (III)	16+
1. Observed wage	1.66	2.25	3.54	14.60	16.33	18.51	4.23	5.52	6.60
2 Capacity wage	1.96	2.59	4.60	17.25	18.87	24.05	5.00	6.41	8.57
3. Depreciation	0.05	0.30	1.40	0.44	2.18	7.32	0.13	0.73	2.63
Sum (2 + 3)	2.01	2.89	6.00	17.69	21.05	31.37	5.13	7.14	11.20
Assuming 1,500 hr work per year							7,740	10,710	16,800
h_c				8.80	7.26	5.23			
N_c				3.45	2.96	2.81			
h_c/N_c				2.55	2.45	1.86			

NOTE.—h_c = duration of home time, N_c = no. of children

Only panel I represents the marginal price of time Note that the observed wage rate[25] represents 80 percent of the marginal price of an hour below college levels and only 60 percent at higher levels The same proportions hold in the other two panels However, figures in these panels are not prices but expenditures which depend on both the price of time and the number of children and the average home-time interval per child Both of these variables can be viewed as responses to the marginal price of time As the table indicates, observed wage rates and, even more so, marginal prices of time (panel I) increase with education Lesser fertility and closer spacing of children are the responses [26] both numbers of children and interval of home time per child diminish Consequently, the differences in total expenditures by education level are reduced While the marginal price of time of the highest education group is three times as high as that of the lowest, the expenditures per child are a little over twice as high, and total expenditures are only 70 percent higher

Since the opportunity costs of labor-force withdrawal ("home time") are not quite the same thing as the opportunity costs of children, we again caution the reader to view the estimates of table 13 as largely illustrative They clearly illustrate the point which the title of this paper intends to convey foregone market-oriented human capital of mothers is a part of the price of acquiring human capital in children, and more generally, a price exacted by family life Of course, the greater market specialization, longer hours, and greater intensity of work and of job training on the part of husbands and fathers can be viewed as a "price exacted by family life" in exactly the same sense

Implicitly, families balance such prices against perceptions of received benefits [27] Of course, both perceptions of net benefits and prices change While perceptions are matters of individual psychology and of cultural climate, the marginal opportunity cost of time has risen secularly with the rise in real wages and with the growth of human capital It is natural for economists to connect to this basic fact both upward trends in labor-force participation of women and downward trends in fertility,[28] changes in the family, and even some of the rhetoric which accompanies these developments

---

[25] In principle, wage rates just before the period *h* are required The wage at ages 35–39 represents, on average, a small overstatement wage profiles of married women with children are relatively flat in the age span 25–39 within education groups

[26] Direct evidence on closer spacing at higher levels of education is shown in research for a Columbia Ph D dissertation by Sue Ross (1973) In the NLS data, there is a strong correlation between the length of home time and the birth interval from oldest to youngest child

[27] Some of these benefits are analyzed in the papers of Lee Benham and Arleen Leibowitz in this volume

[28] For economic analyses which bear on the upward secular trends in labor-force participation of married women, see Mincer (1962) and Cain (1966) For analyses bearing on fertility trends, see "New Economic Approaches to Fertility," *J.P.E*, vol 81, no 2, suppl (March/April 1973)

Appendix

Note on the Construction of Work-Experience Intervals

The 1967 NLS survey of women aged 30–44 permits a division of time elapsed since leaving school into, at most, eight intervals The following information was used in constructing these intervals (*a*) Dates were available for school leaving (S), first marriage (M), birth of first child (C), start of first job, return to labor force after birth of first child, start of current job, and end of last job, if currently not working (*b*) Number of years during which the woman worked at least 6 months between (1) school leaving and first marriage, (2) marriage and birth of first child, (3) return to labor force after the first child, and (4) the start of current job

On this basis, we describe the intervals in the order of their chronological placement interval h_1 (on average, half a year) is the interval between school and first job, e_1 is the number of years of work between school and marriage The placement and continuity of this interval checks rather closely with the data, though direct statements are absent, e_2 is years worked (similarly defined) between first marriage and birth of first child, h_2 is the residual home time, given information on the length of interval between first marriage and birth of first child The assumption of continuity and order of placement of e_2 and h_2 are somewhat arbitrary They are justified by evidence of frequent identity of job e_1 and e_2 and the plausibility of h_2 starting during pregnancy. Indeed, h_2 is a fraction of a year, on average, h_3 is the uninterrupted interval of home time following the birth of the first child. It is placed by direct information, e_3 is years of work and h_4 the residual amount of time in the interval between returning to the labor force at the end of h_3 and start of current job. However, neither e_3 nor h_4 needs to be continuous. The succession of h_4 after e_3 is more plausible than the converse Also ($e_3 + h_4$) is, on average, about 3 years altogether, e_4 is clearly defined and placed as the current job interval.

In tables 2 and 3 we aggregate ($e_1 + e_2$) and call it e_1, ($h_2 + h_3$) is h_2, and the other intervals are correspondingly renamed.

In the regressions we added h_1 to h_3 to get h_2 other home time Separately, or together, these intervals are quite short and show little effect in our analysis

Comment

Otis Dudley Duncan

University of Michigan

For readers, other than economists specializing in studies of human capital, the most controversial sections of Mincer and Polachek's paper are likely to be the introductory statements of the motivation of the research and the authors' calculation which suggests that not less than half the wage gap between married women and men is due to sex differences in length and continuity of work histories.

To their initial rationalization of the division of labor within the family in terms of "complementarity and substitution relations in the household production process" and "absolute and comparative advantages due to differential skills and earning powers," the authors quickly add the disclaimers that sex linkage of intrafamily role differentiation is "subject to technological and cultural changes" and that such changes along with "discriminatory attitudes . . . in the labor market" are outside the scope of the study. Most feminists and many sociologists would presumably respond that with this much of the problem out of scope, what remains is likely to be misleading.

In technical terms, the issue is whether the "statistical residual" u in the general specification of the model is indeed uncorrelated with the vector e so that the coefficients of e_t (experience segments) in the earnings function are unbiased estimates of the investment in human capital appropriate to the wife's role in the household. Mincer and Polachek do address the problem of bias in their somewhat cursory treatment of two-stage least-squares estimation of the earnings function. But there the concern is with possible simultaneity of wage levels and experience variables. The more serious question may be whether such institutional and cultural factors as discrimination and employers' unwillingness to invest in training female workers—all taken to be beyond the scope of the study and presumably captured only in the residual u—act as common causes of wage levels and work experience.

Indeed, in the remarks appended to the calculations on the wage gap, Mincer and Polachek come perilously close to a premature closure of their conceptual scheme. Having attributed half the gap to sex differences in work experience they go on to observe that the lower coefficient for that experience among women may also be attributable to lower work attachment on the women's part. But the sex difference in work experience and the sex difference in coefficient for work experience provide an exhaustive decomposition of the sex difference in wages, net of initial endowment as of the beginning of the work history Hence, the paper seems to be saying that women get about what they deserve in the job market [1]

Many technical and conceptual improvements are required if the strategy opened up by this paper is to be productive Mincer and Polachek seem amply aware of the frailty of their estimates, due to sample size and crudity of measurement The hypercritical reader, however, will not fail to note various places where the authors seemingly read their tables a little too smoothly

On the conceptual side, much work remains to be done before a serviceable model emerges from this kind of work Treatment of the "activities, and characteristics of other family members," mentioned at the outset, is underdeveloped in the present formulation, while the labor-supply function is admittedly inadequate and a fertility function is lacking. More directly to the point of the inquiry at hand, one is disappointed to find no effort to include the quality of the work experience, and this is perhaps a more serious omission than the parallel inattention to quality of schooling One indication that investigation of this facet of the problem could be productive is given by table 4.7 in *Dual Careers* (Parnes, Shea, Spitz, and Zeller 1970, p 114), wherein it appears that the effect of the rate of lifetime labor-force participation on the wage rate is substantially greater for white-collar than for blue-collar workers among currently employed women. In this connection Mincer and Polachek do investigate the interaction of work experience and schooling with respect to wage rate Close inspection of their tables 5 and 6 turns up some disconcerting variations in coefficients by years of schooling At this point errors in the data and misspecifications of the model are hard to tell apart

One can surmise that when all the truly endogenous and relevant causes and effects of the earnings of women are explicitly represented in the model, it will either be underidentified or else have been broadened to include the "sociological" exogenous variables ruled out of scope in this initial and productive foray In trying to get the model properly specified, however, one need not expect too much help from the discipline of sociology as such, geared as it is to the production of dialectic rather than discovery.

---

[1] The authors' own views on the points at issue are made much clearer in their revision which was not available when this comment was prepared.

Home Investments in Children

Arleen Leibowitz

National Bureau of Economic Research and Brown University

By the time children enter first grade, significant differences in verbal and mathematical competence exist among them.[1] These differences reflect variations in (1) inherent ability, and (2) the amounts of human capital acquired before the children reach the age of six.[2] The stocks of acquired human capital reflect, in turn, varying inputs of time and other resources by parents, teachers, siblings, and the child. The process of acquiring preschool human capital is analogous to the acquisition of human capital through schooling or on-the-job training.

Assuming a constant rental rate for human capital, earnings can be interpreted as a measure of capital stocks at later ages. The IQ also can be interpreted as such a measure of human capital stocks. It is related to some commonly used inputs of human capital, for it is well known that measured IQ is not independent of years of schooling acquired before the age of testing. At preschool ages IQ measures should be related to human capital inputs in early childhood as well as to inherent genetic ability.

Viewing measured ability as an index of the stock of human capital puts a different light on earnings functions which include ability and schooling. If contemporaneous ability and schooling measures are used to predict earnings (as in Hansen, Weisbrod, and Scanlon 1970), it is not surprising to find that earnings are more closely related to an ability measure, which

This research has been supported by grants to the National Bureau of Economic Research from the Office of Economic Opportunity and the National Institutes of Health. A National Science Foundation grant to the National Bureau of Economic Research funded the data preparation. I gratefully acknowledge the very competent research assistance of David L. Lindauer and Arun K. Mukhopadhyay. For helpful comments on a previous draft of this paper I thank Charles Betsey, Barry Chiswick, John Hause, Jacob Mincer, and Finis Welch.

[1] This is one of the important findings of the Coleman Report (see Mosteller [1972, p. 49]).

[2] Genetic factors cannot account for the entire difference, as studies of identical twins show. In one such study, identical twins raised apart showed an average difference in IQ of 14 points (Kagan 1969, p. 275).

is an index of human capital stocks, than to years of schooling, which is merely a partial measure of one of the inputs In this respect family-background variables included in an earnings function may be interpreted as proxies for early investments in human capital if characteristics of the father and mother are systematically related to the investments in time and goods that they make in their children

I. A Causal Model

Consider a model in which earnings depend upon human capital stocks which are built up over the life cycle through schooling, on-the-job training, and home investment

By adapting Ben-Porath's well-known model of the acquisition of human capital over the lifetime through investments, we postulate a variant of the Ben-Porath production function for human capital (Ben-Porath 1967, p 360)·

$$Q_t = \beta_0 S_t^{\gamma_1} K_t^{\gamma_2} D_t^{\beta_1}, \qquad (1)$$

where Q_t = gross additions to human capital stock in time t, K_t = th stock of human capital existing at t, S_t = the proportion of K_t allocated t producing Q_t, $0 \leq S_t \leq 1$, D_t = the quantity of goods inputs allocate to producing Q_t, and β_0 = the Hicks neutral-efficiency parameter fo producing human capital, $0 < \gamma_1, \gamma_2, \beta_1 < 1$. The rate of change in th capital stock is given by

$$K_t = Q_t - \delta K_t, \qquad ($$

where δ is a constant depreciation rate The stock at time T can defined as

$$K_T = K_0 + \int_{t=0}^{T} K_t \, dt, \qquad ($$

where K_0 = the initial stock or inherent genetic ability

The marginal-cost curve which corresponds to human capital p duction function (1) is shown by Ben-Porath to be

$$MC_t = \frac{a_0}{\beta_0 \gamma_1} \left(\frac{\gamma_1}{\beta_1} \frac{P_d}{a_0} \right)^{\beta_1/\theta} \left(\frac{Q_t}{\beta_0} \right)^{\theta} \cdot K_t^{(\gamma_1 - \gamma_2)/\theta},$$

where $\theta = (\gamma_1 + \beta_1)$ Thus, the marginal cost of acquiring hun capital is a function of existing stocks if $\gamma_1 \neq \gamma_2$ If $\gamma_1 > \gamma_2$, MC r with the level of existing stocks, and human capital increases the ductivity of time in the market more than the productivity of tim producing new capital. However, if $\gamma_1 < \gamma_2$, the marginal cos producing new capital falls with existing stocks. The "demand pr depends, however, only on the rental price for human capital, the of interest and depreciation, and on the time remaining in the wor

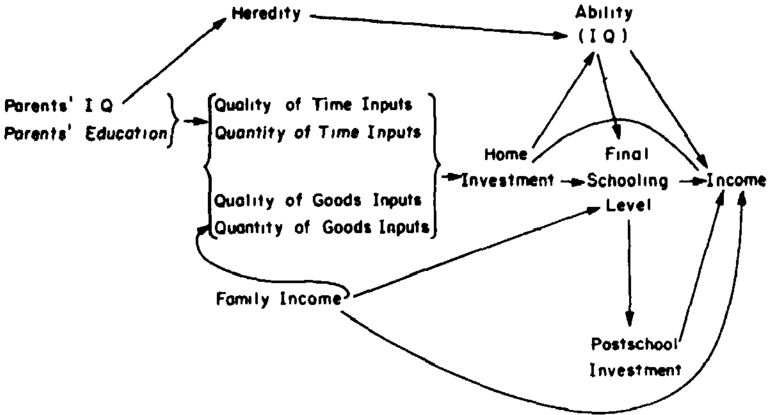


Fig 1

lifetime Thus, we would expect persons with greater initial levels of human capital to demand more additions to the stock if $\gamma_1 < \gamma_2$

Figure 1 helps us to visualize the process Parents' genetic characteristics determine the heredity of the child I have shown (Leibowitz 1972) that the quantity of time devoted to children is positively related to parents' education, and there is evidence that the quality of time inputs is also positively related to education (Schoggen and Schoggen 1968) Thus, parents' genetic endowment and their education determine the quality and quantity of time inputs to the child Parents' attributes also determine family income, which affects the amount and quality of time and goods inputs. These inputs comprise home investment, which, along with heredity, determines IQ Final schooling level is determined by family income and stocks of human capital. Home investments may affect capital stocks in a manner not reflected in measured IQ Thus, a variable such as the education of a mother may affect a schooling level both directly and through heredity Income, which is a rent on the stock of human capital, depends on the four major sources of capital home investment, measured ability, final schooling level, and postschool investment This model can be written as a recursive system

$$
\begin{aligned}
IQ &= f_1(G, I_1), \\
I_2 &= f_2(IQ, I_1, Y_f), \\
Y &= f_3(I_1, I_2, I_3, IQ),
\end{aligned}
$$

where G = genetic factors, I_1 = home investment, I_2 = schooling investment, and I_3 = postschool investment.

Clearly, in this system variables which are endogenous in one equation are predetermined in equations of higher number. This is true because the endogenous variables are determined at various early stages of the life cycle. That is, childhood IQ is determined prior to schooling, which is determined prior to earnings If we assume that disturbances from each pair of equations are uncorrelated in the probability limit, this system may be estimated consistently by the use of ordinary least squares

The rest of this paper reports an investigation of the possible returns to home investment, concentrating on the following questions·

1 Do home investments add to preschool stocks of human capital, as measured by IQ?
2 Is the amount of schooling achieved affected by early stocks of human capital and by later home investments?
3 Do home investments affect earnings if other forms of human capital are held constant? What is the bias to the coefficient on schooling (the rate of return to schooling) if home investments are omitted from the earnings function?
4 Does an early measure of ability affect earnings if schooling and home investments are held constant? What is the bias to the coefficient on schooling if ability is omitted from the earnings function?

To investigate these questions a unique data set called the Terman sample is used, it is described in Section II In Sections III–V, th estimation of the childhood human capital, education, and earning functions is discussed

II. The Nature of the Data[3]

In 1921, Lewis M Terman, a Stanford University psychologist, initiated study to evaluate the physical, mental, and personality traits of Californ school children who scored in the top 1 percent of the national I distribution, and to follow those children as far into adult life as possib to see if high IQ was a good predictor of success in later life (The da generated will be referred to hereafter as the Terman sample.) In 19 when Terman began to test the children, he selected 1,528 of them wi IQ scores of 140 or above, corresponding to the top 1 percent of t intelligence distribution

The original sample consisted of 857 boys and 671 girls The lar proportion of males among the group can probably be accounted for the larger variability in tested IQ among boys than among girls.

---

[3] The data description in this section draws heavily upon discussion in Terman Oden (1959). I am indebted to Susan Crayne of the National Bureau of Econo Research for devising an ingenious method for decoding the column binary tape of data and for doing the decoding

The method of selection was not free from bias. Budget constraints precluded testing every child in the California schools Rather, in grades three through eight of schools in urban areas, each classroom teacher listed the three brightest children in the class and the youngest child These children were further screened by a group IQ test, and the "promising" candidates were given the Stanford-Binet test The original criterion for inclusion in the study group was a Binet IQ score of 140 or above, although 65 pupils who tested between 135 and 139 were also included High school students nominated by their teachers as being bright were chosen if their score on the Terman group test fell within the top 1 percent of the distribution. From later evaluations, it was estimated that 90 percent of the eligible children had been included in the sample The mean IQ of the students on the Stanford-Binet test was 151 5 for boys and 150 4 for girls at the time they were included in the study Average age of Binet-tested pupils was 9 7 years and that of group-tested high school students was 15 2 years at the time they were selected.

The sample is thus not in any sense representative of California school children in the 1920s The children were atypical physically and socially as well as intellectually They tended to be taller and stronger and to mature earlier than their classmates Fathers of the students had a median 12 4 years of schooling and the mothers, 11 7 years—nearly 4 years above the average for their generation in the United States

The achievement of sample members also was above average About 70 percent of the students in the sample finished college as contrasted with 8 percent of the United States population in their cohort The average earnings of males in the sample in 1960 exceeded those of men of equivalent age and years of schooling by 30–50 percent

At the outset, in 1921, in addition to IQ scores, many data about the home backgrounds of the students were collected Information on the test group members was updated in 1929, 1940, 1950, 1954, and 1960 to include current data on schooling, earnings, occupation, and data on other variables Although the data collection spanned almost 40 years, the retention rate in the sample was very high Of the original 1,528[4] sample members, 104 (6 8 percent) died by 1955, after 1928 only 28 were lost track of until, following 1945, six more were not heard from

III. Determinants of Ability

Our discussion up to this point has emphasized that ability as measured by IQ or other similar tests is not only the result of inherited attributes but also of acquired human capital Although schooling level may affect measured ability, this will not be the case in our analysis, where ability is the childhood IQ score determined in 1921–22. The IQ measure is the score on the Stanford-Binet or Terman group test (TGT).

---

[4] Regressions in this paper are based on the 821 males and 643 females for whom complete childhood records were available Sample sizes may be further reduced in certain regressions due to missing data

Several measures of home investment are readily available in the data. The respective educations of the father and the mother are important indices of the quality of the time spent with the child and should be positively related to IQ. In addition, as I have shown elsewhere (Leibowitz 1972), education is also positively related to the quantity of time spent with the child. Since mothers' time expenditures on children exceed those of fathers by at least a factor of 4,[5] we would expect the significance and size of the coefficient of mother's education to exceed that of father's education. However, education of the parents may be a proxy for the quality of genetic inheritance—if intelligence of the parents and their own education are positively correlated, and if intelligence can be transmitted genetically. If education is acting as a proxy for heredity, the coefficient of each parent's education on child's IQ should not differ significantly. A comparison of the size and significance of these two variables will indicate whether parents' schooling is influencing IQ directly through the quality and quantity of time spent with children or whether it is a proxy for inherited ability.

A more direct measure of the quantity of time spent with children in preschool investment results from the parents' estimate in 1922 of how much home instruction they provided their child. Two dummy variables were generated: HOMETR2 = one if parents conducted "an appreciable amount of instruction along particular lines," and zero if they did not; HOMETR1 = one if parents spent a "considerable number of hours, but chiefly reading, telling stories, and writing," and zero if they did not do so. The omitted category covered "no instruction, other than usual amount of reading and telling stories."

It is a consistent finding of psychological studies that first children and children from smaller families score higher on IQ tests and achieve more (for instance, get more schooling). The analysis of family investment presents a rationale for this phenomenon, since the more children there are, the less time input parents have available per child. For the same reason children of lower birth order can be expected to receive more time inputs. Both the number of siblings and birth order have been included as proxies for home investments.

In addition, the regressions included a dummy variable equal to one if the IQ test was the Terman group test rather than the Stanford-Bine test. About 33 percent (268) of the male subjects and 26 percent (168) o the females had taken the TGT, in which average scores were 8.5 point below the Stanford-Binet version of the IQ test. Terman's analysis showe the TGT-tested students were at least as select a group as the Binet-tested ones, but scores were shifted down The assumption is made that the two tests are identical in all other respects Means and standard deviations of the variables are presented in Appendix C

---

[5] This is based on 1967 data (Leibowitz 1972). It appears unlikely that in 1922 fathe spent more time with children than mothers did.

A. Estimation for Males

In table 1 the regression of family-investment variables on IQ measured at age 11 is presented for the 821 males in the sample Both home-investment variables were positive, with HOMETR2 having a stronger positive impact on IQ than HOMETR1. Thus, the quality of time inputs by parents did raise measured IQ and special instruction had more of an impact than generalized time inputs Since this holds the quantity of time inputs constant, the education variables can be taken to measure the quality of time The mother's education is a very significant predictor of IQ, while the coefficient for father's education is not significantly different from zero This suggests, as argued above, that education is a proxy for home investment and not solely for inherited factors Psychologists have also shown (Kagan and Moss 1959) that correlation between child's IQ and maternal education is greater than that between child's IQ and paternal education, while the correlations between child's IQ and those of mother and father do not differ significantly

Birth order is represented by the variable BORDER, which was coded 1 for a first-born child, 2 for the second-born, and so on This variable thus contains some information about number of siblings as well as about birth order The negative effect of birth order, while quantitatively small, is quite significant If, instead of birth order, number of siblings is used, its coefficient is significantly different from zero only at the 10 percent level of significance If both are included in the equation, birth order's coefficient remains significantly different from zero, while number of siblings does not Estimated family income in 1922 (YK) did not prove to be a significant predictor of childhood IQ, as seen in table 1, column 2

When the equations are based on the 553 boys who took the Stanford-Binet test, thus allowing us to dispense with the TGT dummy variable, the coefficients are substantially the same (see col. 3, table 1)

B Estimation for Females

The same variables which accounted for 18 6 percent of the variance in IQ of males in the sample account for only 13 percent of the variance of females' IQ (see table 1, col 4) Apart from the TGT dummy and family income, none of the variables meets conventional levels of significance Neither the direct measures of the quantity of time inputs (HOMETR1 and HOMETR2) nor the parents' education variables have coefficients

TABLE 1
DETERMINANTS OF CHILDHOOD IQ

INDEPENDENT VARIABLE	MALES			FEMALES				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TGT	−8 468	−8 498		−7 82	−7 77	−7 80		
	(−12 23)	(−12 19)		(9 30)	(9 28)	(9 29)		
EDMOT	0 3475	0 332	0 420	0 160		0 163	0 115	0 317
	(2 42)	(2 29)	(2 03)	(0.98)		(1 00)	(0 52)	(2 13)
BORDER	−0 561	−0 568	−0 689	0 0008			−0 055	0 299
	(−2 13)	(−2 16)	(−1 92)	(0)			(0 14)	(1 08)
HOMETR2	2 400	2 42	2 86	−0 958			−2 00	2 14
	(2 61)	(2 63)	(2 21)	(0 86)			(1 37)	(1 93)
HOMETR1	1.593	1 593	1 654	0 431			0 344	0 443
	(1 89)	(1 89)	(1 40)	(0 44)			(0 26)	(0 50)
EDFAT	0 065	0 008	0 128	−0 123		−0 135	−0.103	−0 124
	(0 60)	(0 05)	(0 81)	(0 85)		(0 94)	(0.53)	(1 09)
YK		3 79		10 85	9 57	10 97	13 96	0 002
		(0 80)		(2 01)	(2 32)	(2.04)	(1 98)	(0)
C	147 1	144 6	145 6	140 4	141 9	140 3	138 2	139 3
R^2	186	186	040	130	126	129	017	.047
Number of observations	821	821	553	643	643	643	475	168

NOTE.—*t*-values are in parentheses.

significantly different from zero. Birth order and number of siblings have no significant impact on measured IQ. The only other variable which was significantly related to IQ was YK, a measure of relative family incomes, generated from data on the father's occupation in 1922. (The construction of this latter variable is described in Appendix A.)

When the equation is estimated for only those subjects who took the Stanford-Binet test, the only significant variable is family income (table 1, col. 7). However, when the sample covers only the girls who took the Terman group test (as in table 1, col. 8), mother's education and HOMETR2 are positively related to childhood IQ as is the case for males, while family income is not significantly related to IQ. It should be noted that the girls chosen by the TGT were older, and probably constituted a more selective group than the girls chosen by the Stanford-Binet test. The ratio of boys to girls in the TGT group was 1.83 and in the Stanford-Binet group 1.16 (Terman 1925, p. 560). The greater selectivity of the older group is consistent with the fact that IQ of girls falls more than that of boys over time (Sontag, Baker, and Nelson 1958, pp. 22–32). The TGT group had high abilities which endured at least until their adolescence.

Although a drop in IQ on retest is expected on statistical grounds due to regression toward the mean, girls showed greater decreases in IQ than boys when their IQs were retested after an interval of several years. In this sample, 54 children younger than age 13 in 1927–28 were again given the Stanford-Binet test. The 13-point average decrease in IQ for girls was significantly different from the three-point average drop for boys.[6]

These results suggest that the kind of intelligence that endures through adolescence (as was the case with the TGT-selected girls) is systematically related to home investments. However, in the younger group the factors which account for their scoring in the top 1 percent of the IQ distribution are more evanescent and are not related to home investments. Home investments were related to older girls' IQ differences that were stable over time, but only income was related to younger girls' IQs which faded over time.

IV. Determinants of Schooling Level

The model postulates that final schooling level depends on ability, family income, and home investments. If human capital is not equally productive in the market and in the acquisition of new human capital, measures of human capital stock (such as childhood IQ) should be related to years of schooling.

---

[6] The approximate standard deviation of the difference is two (Terman, Burks, and Jensen 1930, p. 25). Four hundred older students also were retested in 1927–28, but the difference by sex was not significant (ibid., p. 33).

The impact of family income on schooling attainment may be quite attenuated in this sample. First, all sample members were residents of California and had access to the "free" state university system, which is of high quality. Second, because of their demonstrated scholastic aptitude, the students without sufficient family resources to finance the indirect costs of schooling would have been eligible for scholarships. Two alternative estimates of parental income were constructed (as described in Appendix A) Further, additional numbers of siblings in a family are assumed to have the effect of reducing the amount of family support for schooling and to indicate lower levels of home investments. The effects of the quantity of preschool home investments (the HOMETR variables) should have been captured in the childhood IQ measure, but measures of home investments made after the childhood IQ test was taken can be expected to have a positive impact on schooling levels Later home investments may be assumed to be positively correlated with early home investments as well as with parents' education If the constructed income variable measures family income imperfectly, father's education may be a proxy for income as well as for home investment (Definitions of schooling levels and their frequencies are given in Appendix B)

A. *Estimation for Males*

Table 2 presents the equations estimating education levels for the 780 males for whom we had 1940 data and the 781 males for whom we had complete 1950 data

The regressions for both years show that childhood ability is positively related to schooling level This fact indicates that the kind of human capital that IQ represents enhances productivity in acquiring schooling more than it does productivity in the labor market This is consistent with the general finding of psychologists that the one thing IQ predicts best is success in school

Both the education of the mother and of the father are positively and significantly related to schooling achieved, and the coefficients are no significantly different from each other The effect of other children in th family (SIBS) is negative and significant at the 10 percent level in 1940 By 1950, the subjects were nearly 40 years old, and the effect of siblings i restricting resources available for schooling seems to have leveled off a other sources of funding were developed

The quantity-of-home-investment variables are not significantl different from zero at the 10 percent level. Since these variables we important determinants of IQ their negative signs may indicate th intensive training may temporarily boost IQ (as measured at age 11 but that this increase fades as time passes. This, the quantity-of-hom investment variables are correlated with a transitory increase in I

TABLE 2
YEARS OF SCHOOLING OF MALES IN TERMAN SAMPLE

	1940		1950	
INDEPENDENT VARIABLE	(1)	(2)	(3)	(4)
EDMOT	0 124	0 131	0 134	0 141
	(3 05)	(3 18)	(3 35)	(3 47)
EDFAT	0 110	0 122	0 149	0 163
	(3 51)	(3 22)	(4 82)	(4 35)
CIQ	0 021	0 023	0 024	0 025
	(2 16)	(2 29)	(2 46)	(2 57)
TGT	0 753	0 765	0 500	0 511
	(3 54)	(3 59)	(2 38)	(2 43)
HOMETR1		−0 192		−0 124
		(0 80)		(0 53)
HOMETR2		−0 293		−0 248
		(1 13)		(0 96)
SIBS	−0 118	−0 123	−0 058	−0 063
	(2 01)	(2 09)	(1 01)	(1 08)
YK		−0 782		−0 929
		(0 58)		(0 69)
C	10 03	10 38	9 39	9 85
R^2	078	080	102	104
Number of observations	780	780	781	781

NOTE —*t*-values are in parentheses

which has faded by the time schooling decisions are made[7] Family income was not significantly related to schooling levels

In summary, these regressions indicate that the schooling level for males is positively related to ability, that larger numbers of siblings may cause schooling to take longer to complete due to the difficulty of financing it (or the obligation to put younger brothers and sisters through school), but the final level of schooling is independent of family size Finally, the levels of parent's education, representing the quality and quantity of home investments, have a positive effect on schooling levels, while the quantity of preschool investments does not Mother's and father's education have equal impacts on schooling, while mother's education was much more important in determining IQ Father's education may be a proxy for income, although a more direct measure of income was not significantly related to schooling levels

B Estimation for Females

In contrast to males, schooling levels of females are not significantly related to ability (see table 3, cols 1 and 4) They are, however, related

---

[7] This transitory effect on IQ of intensive preschool programs has also been seen in programs outside the home, such as Head Start Finis Welch pointed out this analogy to me.

TABLE 3
YEARS OF SCHOOLING OF FEMALES IN TERMAN SAMPLE

INDEPENDENT VARIABLE	1940 (1)	1940 (2)	1940 (3)	1950 (4)	1950 (5)	1950 (6)
EDMOT	0.163	0.180	0.161	0.158	0.180	0.152
	(4.67)	(2.89)	(3.77)	(4.73)	(3.01)	(3.70)
EDFAT	0.085	0.085	0.112	0.073	0.043	0.113
	(2.73)	(1.52)	(3.55)	(2.43)	(0.79)	(3.70)
CIQ	0.011	−0.011	0.014	0.009	−0.019	0.012
	(1.28)	(0.36)	(1.51)	(1.15)	(0.63)	(1.44)
TGT	0.669			0.491		
	(3.47)			(2.65)		
SIBS	−0.136	−0.152	−0.146	−0.126	−0.131	−0.143
	(2.42)	(1.62)	(2.08)	(2.34)	(1.47)	(2.12)
YK	1.524	0.682	0.076	1.705	1.034	0.083
	(1.30)	(0.31)	(1.30)	(1.51)	(0.48)	(1.44)
FIRSTBORN	0.349	0.513	0.274	0.232	0.340	0.164
	(1.97)	(1.60)	(1.26)	(1.36)	(1.10)	(0.79)
C	9.311	13.637	10.004	9.817	14.978	10.531
R^2	.166	.177	.153	.154	.145	.152
Number of observations	606	158	448	615	158	457

NOTE.—*t*-values are in parentheses.

to parents' schooling levels. Although the difference between the coefficients on mothers' and fathers' education is greater for the girls than for the boys, this difference is not significant at the 5 percent level. As with males, the more children in the family, the less schooling achieved, but for women the negative effect of a larger family on schooling attainment is not overcome by 1950. Girls who were the oldest child in their family got more schooling, as seen by the positive effect of the variable FIRSTBORN. For boys, once number of siblings was held constant, birth order did not matter. Family income and quantity of home investment in preschool years are not strongly related to years of schooling achieved.

In table 3, columns 2 and 5 refer to the girls selected in high school by the TGT and columns 3 and 6 refer to the Stanford-Binet-selected girls. The mother's education seems to be relatively more important in determining schooling attainment of TGT-girls, while father's education, family income, and number of siblings are relatively more important determinants of schooling for the Stanford-Binet group.

In summary, there is some evidence here that a mother's characteristic have a stronger impact on schooling achieved by females than by males particularly for the girls chosen in high school. Further, number o siblings and family income are more closely related to females' schoolin levels than to those of males, while ability is more strongly related to th males' schooling levels than to those of females. If we apply Becker analysis (1967, p. 17), these results suggest that for girls the supply curv of funds shifts relative to a comparatively stable demand curve fc schooling, while the opposite is true for males. This implies that males face greater equality of opportunity for schooling than do females

C Total Effect of Mother's Education

The total effect of mother's education on schooling achieved is composed of the direct effect and the indirect effect via IQ For a male, each year of mother's education results in an addition of 14 of a year of schooling in 1940 and 15 of a year in 1950 Thus, the effect of having a mother who is a college graduate compared to one who is a high school graduate is to increase schooling attainment by 6 of a year

For girls who took the TGT, each additional year of a mother's schooling resulted in 18 of a year of schooling attained in 1940 and 17 of a year in 1950 For the girls who were selected through the Stanford-Binet test, an additional year of mother's schooling added 16 of a year to schooling attained by 1940 and 15 of a year to schooling attained by 1950

V. Determinants of Income

A Schooling and Postschool Investments

According to the model presented in figure 1, income is derived from the rent on different forms of human capital ability and home, school, and postschooling investment The relation to be estimated is of the form $Y = \beta_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 (I_3)^2 + u$, where Y = log of earnings in a given year, I_2 = years of schooling, I_3 = years of experience, and u = other income-affecting factors which are uncorrelated with the investment variables

Using the approximation that $\log (1 + r) \cong r$, and assuming that rates of return are constant over time, the coefficient β_2 can be interpreted as equal to rK, where r is the rate of return to investments made in schooling and K is the ratio of investment costs to gross potential earnings during the schooling period [8]

A common assumption in interpreting human capital earnings functions is that $K = 1$, since opportunity costs probably account for the major part of investment costs and part-time earnings are assumed to offset tuition charges Under these conditions, β_2 can be interpreted as the rate of return to schooling The inclusion of the number of years of experience and its square results from a Taylor series expansion of a series of on-the-job investments whose ratio to earnings declines linearly with time (see Mincer 1973, pp 3–6).

The parameters of the above equation were estimated using data on income earned by men in the sample in 1939, 1949, and 1959 Since labor-supply data were never obtained, the earnings functions could not be estimated for women in the Terman sample. Only males who were no longer students and who were not unemployed were included in order to achieve as nearly as possible a sample of men who were employed full time.

---

[8] This approach is fully developed by Becker and Chiswick (1966)

The dependent variable in 1939 was the full-time monthly wage or salary. For 1949 and 1959, the dependent variable was annual earned income Years of schooling were assigned to correspond to rather detailed descriptions of schooling level given in the questionnaires (see Appendix B). Experience was calculated as: age − age at high school graduation − (final education level − 12) If we employ the useful assumption that $K = 1$, the regression results presented in column 1 of tables 4–6 imply a rate of return to schooling of between 5.6 and 6.7 percent This is low compared with other estimates of rates of return to college graduates

Rates of return calculated for graduate training are usually lower than those for bachelor of arts degree holders, and 44 percent of this sample had more than 16 years of schooling. There is also reason to believe that the assumption that $K = 1$ is less tenable for this sample than for the average college graduate, since 24 percent of the Terman undergraduate had scholarships or fellowships amounting to $200 or more and 10 percen had received $1,000 or more in support of their studies If $K < 1$ $\beta_2 < r$, thus coefficients on years of schooling appear to be underestimate of the rate of return to education for this sample.

TABLE 4
EARNINGS FUNCTIONS FOR MALES IN 1940

	Log of Monthly Earnings ($)			
	(1)	(1a)	(2)	(3)
Years of schooling in 1940	0.056 (5.18)	0.056 (5.18)	0.063 (5.77)	0 (5
Years of experience in 1940	0.004 (0.07)	0.019 (2.70)	0.019 (2.75)	0 (2.
(Years of experience)^2	0.0003 (0.27)			
Mother's education			−0.016 (−1.52)	
Father's education			−0.009 (−0.91)	
YK			0.85 (2.41)	
Family income adequacy			0.050 (3.40)	
CIQ				− (−
TGT				− (−
Intercept	4.424	4.247	3.529	
R^2	.036	.036	.063	

SOURCE —Calculated from the Terman sample, includes 724 observations
NOTE —t-values are in parentheses

TABLE 5
EARNINGS FUNCTIONS FOR MALES IN 1950

	LOG OF ANNUAL EARNINGS ($100)		
	(1)	(2)	(3)
Years of schooling in 1950	0.062	0.064	0.064
	(6.14)	(6.29)	(5.79)
Years of experience in 1950	0.131	0.138	0.128
	(2.39)	(2.52)	(2.34)
(Years of experience)^2	−0.0018	−0.0019	−0.0017
	(−2.19)	(−2.32)	(−2.06)
Mother's education		−0.026	
		(−2.56)	
Father's education		0.00075	
		(0.08)	
YK		0.655	
		(2.00)	
CIQ			0.00049
			(0.20)
TGT			−0.033
			(−0.52)
Intercept	1.02	0.567	0.905
R^2	.061	.073	.062

SOURCE.—Calculated from the Terman sample, includes 731 observations.
NOTE.—*t*-values are in parentheses.

TABLE 6
EARNINGS FUNCTIONS FOR MALES IN 1960

	LOG OF ANNUAL EARNINGS ($100)		
	(1)	(2)	(3)
Years of schooling in 1950	0.067	0.067	0.075
	(6.81)	(6.63)	(6.93)
Years of experience in 1960	0.097	0.104	0.089
	(1.44)	(1.54)	(1.31)
(Years of experience)^2	−0.001	−0.001	−0.001
	(−1.45)	(−1.54)	(−1.18)
Mother's education		−0.016	
		(−1.58)	
Father's education		0.001	
		(1.05)	
YK		0.263	
		(0.84)	
CIQ			0.002
			(0.86)
TGT			−0.096
			(−1.54)
Intercept	1.722	1.411	1.33
R^2	.091	.097	.097

SOURCE.—Calculated from the Terman sample, includes 701 observations.
NOTE.—*t*-values are in parentheses.

There is, however, a striking constancy in the schooling coefficients from decade to decade. The experience variables are not significant in explaining 1940 incomes (the subjects were, on the average, 29 years old). At this stage in the life cycle, experience may be exerting a linear rather than parabolic effect on earnings, as indicated in column 1a, table 4.

B Home Investments

What is the effect of expanding this earnings function to 'include home investment variables? The equation to be estimated is of the form

$$Y = \beta_1' + \beta_2' I_2 + \beta_3' I_3 + \beta_4' (I_3)^2 + \beta_5 I_1 + V, \qquad (5$$

where I_1 = the amount of home investment and all other variables ar defined as above

It is clear that if home investments affect income and if home investmer and schooling are positively correlated, then the simple regression income on schooling results in a coefficient on schooling which is biase upward from β_2' by a factor equal to $\beta_5 r_{I_1,I_2}$, where $r_{i,j}$ is the regressio coefficient of variable i on variable j in the sample However, if hon investments do not directly influence income, but affect it only indirect by influencing schooling levels, then the coefficient β_2 is not biased u ward from β_2'

Estimates of equation (5) for 1940, 1950, and 1960 are presented column 2 of tables 4–6. Father's education was not significantly relat to earnings in any one of the three years. Mother's education v negatively related to earnings in all 3 years, and in 1950 this coeffici was significantly different from zero Mother's education was positiv correlated with the two family-income measures (correlation w YK = 43, correlation with family income adequacy = .16). Both fam income measures were quite significantly related to 1940 income. 1950, only YK was significantly related to income, and by 1960 neit income measure was Thus, the effect of parents' income on earning attenuated over the life cycle. Family-income variables are relate quality of college attended, and this may, in turn, affect earnings Wachtel 1973) Of those who had "abundant incomes," 63 per attended Stanford; in contrast, only 27 percent of those with "limi incomes and 36 percent of those with "adequate" incomes did so. Gre percentages of the lower-income groups attended state colleges rather the higher-quality state universities.

Mother's education and YK are positively correlated with own ye schooling ($\rho = .23$, $\rho = .17$, respectively) and FAMINC is negat related ($\rho = -.02$). The net effect of the home-background variab to increase the coefficient on years of schooling from .056 to .063 in although the YK variable is significantly related to 1940 income The years-of-schooling coefficient is virtually unchanged by the inclusion of background variables in 1950 and 1960.

These findings provide some evidence against the hypothesis advanced by Bowles (1972) that specification errors and errors in measurement of family-background variables lead (because these variables are positively correlated with schooling level) first to overestimating the importance of schooling; and second, to underestimating the importance of background variables in the earnings function

Bowles maintains that because data on parents' education, occupations, and incomes are obtained from questionnaires administered many years after the fact "the degree of error in the measurement of the father's occupation and education variable greatly exceeds that in the respondent's own years-of-schooling variable" (1972, p S227) Consequently, the importance of background variables will be underestimated, while that of the respondent's own schooling will be overestimated The Terman data allow a simple test of the proposition that if background and the sample member's own data were subject to similar levels of error, the background variables would increase in significance while the member's own variables would fall In the Terman sample, data on parents' education and home investment were supplied by parents when the child was 11 years old Yet, in spite of the significance of some of these variables, the importance of the years-of-schooling variable was not reduced.

C. *Ability*

Using the same reasoning applied to the home-investment variables, we can say that if ability affects earnings and if it is positively correlated with schooling achieved, omitting this variable in the earnings function results in biasing upward the rate of return to schooling.

Income in 1940, 1950, and 1960 is not significantly related to IQ, holding schooling and experience constant (see col 3 of tables 4–6). The IQ variables remain insignificant if both home investment and IQ variables are included in the earnings function, while the coefficients on the schooling and experience terms are virtually unchanged. We conclude that childhood ability has very little independent effect on earnings It is to be expected that earnings are related to contemporaneous stocks of human capital. In this sample, years of schooling in 1940 is no doubt a better index of human capital in 1940 than is IQ measured in 1922

Furthermore, it must be recalled that due to the nature of the sample, we are considering marginal differences in ability within the top 1 percent of the ability distribution Yet, IQ scores varied for this group from 135 to 200, while a standard deviation for the IQ distribution is 15 points and IQ was a significant determinant of schooling.

With a less restricted sample, Griliches and Mason (1972) who used data on United States military veterans who were younger than 35 years, also found the inclusion of an ability measure did not greatly change the coefficient of schooling. Using the change in schooling level after military service, they concluded that the bias in this coefficient due to the omission of an ability variable (AFQT) was only about 12 percent. In their sample, in contrast with this study, the ability variables were significant, although they did not greatly change the schooling coefficient.

Several other studies have also found relatively small bias due to omission of ability variables. Ashenfelter and Mooney (1968, p 86), in a study of past recipients of Woodrow Wilson fellowships, found that the "inclusion of an ability measure affected the estimates of the coefficients for the other education related variables only in a very marginal fashion " Weisbrod and Karpoff (1968), roughly controlling for ability in a sample of employees of the American Telephone and Telegraph Company, estimate that only one-fourth of the earnings' differences between high school and college graduates is attributable to nonschooling factors rather than to differences in schooling level.

Taubman and Wales (1972) found, after testing several ability measures that in the NBER-Thorndike sample of United States Air Force volunteer
for certain programs, only the omission of mathematical ability affecte
the schooling parameters They estimate that the bias to the educatio
coefficient from omitting the mathematical-ability variable was abou
25 percent for 1955 incomes and 15 percent for 1969 incomes Usin
unpublished data from Dael Wolfle on Minnesota high school graduate
Taubman and Wales found a bias of "no more than four per cent at t
various educational levels" from omitting ability measures from t
earnings function.

John Hause (1972) found that the effect of ability on log of inco
remained small at all schooling levels when he used data from Proj
Talent which tested a sample of high school students and later survey
their earnings He calculated upward bias in the education coefficient
ability is omitted, at 13–18 percent in Daniel Rogers's (1969) data ba
on a sample of Connecticut school boys, and at 3–11 percent in
Thorndike data.

VI. Summary

Because of the unique nature of the sample, the findings on the Ter
data cannot be readily generalized to the population as a whole. Howe
several conclusions can be drawn:

1. It has been demonstrated that home investments do incr
measured stocks of childhood human capital. Even within a sampl

very able children, home-investment variables were positively and significantly related to a measure of human capital, IQ, for boys and for a subset of older girls. A mother's education was significantly related to IQ, while the father's was not, thus indicating that home investments rather than wholly genetic factors underlie the relationship.

2 Education achieved by age 29 and age 39 by men and women in the Terman sample did depend on parents' education and on family size, but not on a more direct measure of the quantity of preschool time inputs or on family income. Boys' achievement also depended on ability, indicating greater equality of opportunity for boys to get schooling than for girls.

3. Men's earnings at ages 29, 39, and 49 were strongly related to schooling and experience variables. The addition of home-investment and ability variables did not significantly decrease the coefficients on schooling and experience, although family income was positively and significantly related to earnings in the early years of the life cycle.

Appendix A

Measures of Family Income

Two measures of family income were examined in this study. One was a subjective measure, the other was constructed from data on fathers' occupations in 1922.

A The subjective measure, family income adequacy, resulted from the answer given in 1950 to the following question "How adequate were family finances during your childhood and youth?" Answers were coded as follows

Coding	Frequency (in %)
1 = limited or very limited	20.9
2 = adequate	75.0
3 = abundant	4.1

B An alternative estimate of 1922 family income was derived from the father's 1922 occupation, reported as one of six categories. Estimates of income for various occupation groups were derived from King (1923, p. 111) since the United States Census did not collect income data by occupation until 1940 King's estimates were based on average earnings for a sample of 436 employed males between December and February 1922.

The ratio of income in each category to the income of professional workers is presented below

	Terman Category	YK
1	Professional	1.00
2	Semiprofessional, higher business	0.90
3	Clerical, skilled trades, retail business	0.82
4	Farming and agriculture	0.78
5	Semiskilled trades, minor clerical, or business	0.72
6	Slightly skilled trades, little training	0.64

Appendix B

TABLE B1

YEARS OF SCHOOLING IN THE TERMAN SAMPLE, DESCRIPTION, AND 1950 FREQUENCIES FOR MALES AND FEMALES

YEARS OF SCHOOL CODED	HIGHEST LEVEL OF SCHOOLING	MALES		FEMALES	
		Number	%	Number	%
10	High school, 1–3 years	7	0 9	9	1.4
12	High school, 3 years plus business or trade school, high school graduate	59	7 2	63	9 8
13	High school graduate plus 1–2 years of art, music, business, normal, technical, or trade school, or less than 2 years of college	65	7 9	54	8 4
14	High school graduate plus pharmacy, chiropractic, accounting, or nursing school or 2½–3 years of college	86	10 5	70	10 9
15	High school plus 3–4 years of art or music school or college	5	0 6	1	0 2
16	B A or L L B or D D S without B A	200	24 3	186	28 9
17	Masters' degrees or 5–6 year engineering or architecture degrees	138	16 8	213	33 1
18	D D S with B.A , M B A., graduate work of 2 years	20	2 4	1	0
19	J D or L L B with B A	87	10 6	2	0
20	Ph D , Ed D , L L D , etc	74	9 0	13	2
21	M D	45	5 5	6	0
..	Incomplete information	36	4 4	25	3

Appendix C

TABLE C1
Means and Standard Deviations—Males

Variable	Mean	Standard Deviation
A 1922 data—821 observations:		
CIQ	148.88	10.36
EDMOT	11.74	2.71
EDFAT	12.36	3.53
SIBS	1.85	1.58
ORIGTEST	0.33	0.47
BORDER	1.89	1.24
HOMETR1	0.19	0.40
HOMETR2	0.15	0.36
YK	0.88	0.09
FAMINC	1.82	0.48
B 1940 data—724 observations		
LWAGE40	5.60	0.67
ED40	16.11	2.65
EX40	8.82	3.90
EX40SQ	92.92	75.63
YK	0.88	0.09
FAMINC	1.81	0.48
C 1950 data—742 observations		
LINC50	4.30	0.81
ED50	16.38	2.59
EX50	18.18	4.27
EX50SQ	348.69	155.79
D 1960 data—699 observations		
LINC60	4.91	0.60
EX60	28.16	4.31
EX60SQ	811.51	243.27

TABLE C2
Means and Standard Deviations—Females

Variable	Mean	Standard Deviation
A 1922 data—643 observations:		
CIQ	148.33	9.96
EDMOT	11.88	2.67
EDFAT	12.09	3.55
SIBS	1.88	1.60
ORIGTEST	0.26	0.44
BORDER	1.98	1.33
HOMETR1	0.19	0.39
HOMETR2	0.13	0.34
YK	0.89	0.09
FAMINC	1.83	0.51
B. 1940 data—606 observations:		
ED40	15.34	2.12
C 1950 data—615 observations		
ED50	15.50	2.03

Comment

Frank P. Stafford

University of Michigan

The Leibowitz study is important because, through the use of hom investment variables, it goes beyond the usual practice of recording th parental education and income should in some fashion influence th lifetime-earnings potentials of individuals As the study suggests, da requirements for research on home investments are rather demandi because such final output measures as adult earnings are not oft available along with measures of home and school investments Moreov the position advanced and given considerable empirical support is th the family investments end up as being a type of intermediate good. T intermediate-good argument is crucial to the analysis, for it suggests th simply taking measures of human capital for adults (education and lab market experience) and tacking on variables for home investments n suggest only a minor influence of the latter which would likely be s sumed in measures of adult skills

While the author has demonstrated that home investments substanti affect adult earnings, there are several theoretical and empirical probl with the current effort which may be remedied in future research.

It would seem that to pursue seriously the specification of a dyna accumulation model starting at childhood (even given an exoge family size and earnings potential of the parents) would require somet more than Ben-Porath's model of investment in human capital. F K_0 is the initial condition of "inherent genetic ability," it seems unl that preschoolers choose, or that parents can make them choose, a path of training so as to maximize the present value of net earnings. large extent, the investments in preschoolers are organized by parents, and this preschool investment determines an intermediate This intermediate good then contributes to the learning process point when the child begins to make independent investment de with his parents' resources defining a financial constraint or margin of funds. A variety of specifications, such as having the intermediate good an argument in the production function or altering the efficiency parameter, could be used. This general specification would be more in keeping with the empirical work in the paper and would separate conceptually the role of parents and children in determining educational attainment

A more difficult theoretical problem is hidden in the recursive model used. While a model of this type has much to recommend it on grounds of estimation, the critical factors influencing home investment in children include number of children in the family and the quantity and quality of home investments as well as other training given to particular children. That is, the process of investment in children is bound up with the family decision on number of children. On a descriptive level it may be sufficient to treat the number of children as exogenous and observe how variations in parental input influence educational attainment, given family income Yet, the decision making by the family requires a delicate balancing of number of children and lifetime family income by virtue of the very high (dynamic) shadow price of the wife's withdrawal from the labor force, particularly for highly educated young women Although these decisions can be portrayed by a complex model, there is much merit in a less ambitious (realistic?) approach which takes the family size and home investments per child as given The point of reference to a model with a demographic sector is to emphasize that the approach in the Leibowitz paper must be utilized with some uneasiness insofar as it glosses over a large and important question of family behavior How do parents decide number of children and home inputs per child?

On the empirical side, the major data limitations are that in the Terman sample there is only minimal variation in several of the critical variables (IQ has a mean of 149 with a standard deviation of only 10) and that the variables measuring home training, HOMETR1 and HOMETR2, are rather vague. It seems tenuous to argue that the difference between (1) "no instruction, other than usual amount of reading and telling stories" and (2) "considerable number of hours, but chiefly reading, telling stories and writing" or between (1) and (3) "an appreciable amount of instruction along particular lines" really constitutes a measure of the quantity of time inputs by the parents. Yet, this is in fact assumed because use of these variables, it is argued, "holds the quantity of time inputs constant, [and] the education variables can be taken to measure the quality of time." This is all the more difficult to believe since recent work (Leibowitz 1972, Hill and Stafford 1973) strongly suggests that quantity of time inputs to children by more-educated women are greater than for less-educated women within a lifetime income group.

Without distinguishing between quality or quantity of home inputs per child, the finding, verified in other studies, that education of the wife is much more important than education of the husband (or IQ of the

mother) in determining IQ of the child implies that there is a larg
nongenetic component to IQ. More significantly, this nongenetic com
ponent which can be related to home investment influences education
attainment This is a first major step in the direction of showing th
home investments constitute an intermediate form of human capit
which influences education. The second step is the demonstration th
the IQ and parental-education variables (home investment proxie
influence educational attainment. The third step is to show that educati
influences earnings but that the home investment variables have no dire
influence on earnings This finding corroborates recent work for
national probability sample wherein each year of a father's educati
results in a .3 increase in the years of schooling of his sons, but the earni
of sons are not directly influenced by education of the father (John
and Stafford 1973, pp. 145 and 151)

A secondary theme developed in the paper is the issue of upward b
in educational variables owing to measurement error attributable
imperfect memory or subjective recall in background variables W
minor error in the background variables because the information
obtained from parents when the child was 11 years old, it is found that
influence of the child's own schooling is not appreciably lowered T
results are interpreted as contradicting the contention that most rese
understates the importance of background variables (Bowles 1972)
aside from the issue of measurement errors, the model used here spe
that parental variables influence home investment and therefore ha
major, though indirect, influence on adult earnings This and
research (Griliches and Mason 1972, Johnson and Stafford 1
constitute an empirical link in understanding the rule of intergenerat
influences on the income distribution.

To conclude, let me mention two topics on which this study
bearing The observed positive correlation between spouses' educ
can be interpreted as a mutual consumption choice and the retu
education of women is, under this view, partly realized in this fa
(though Veblen interpreted this consumption as a vestigial form of
atory activity) In the Leibowitz study, one of the major retu
education of women is in the form of home investments in childre
interesting study would be to develop value-added figures and co
these with earnings data to calculate returns to schooling for wome
would show, I believe, that earnings differentials between educate
and women are considerably offset by nonmarket productivity diffe

The efficacy of public policy as a vehicle for altering the
distribution is limited since, it is said, income is so strongly influen
parental background that opening better schooling opportunities
have any appreciable effect on earnings for those born into low-
families. In this light, do the different forms of income maintena

allowing more home time by the mother, result in larger home investments in the children? If the answer is yes, there is reason to believe that public policy can influence what appears to be an important source of human capital formation

Effects of Child-Care Programs on Women's Work Effort

James J. Heckman

Columbia University and National Bureau of Economic Research

In recent years, Congress has considered a variety of work-subsidy grams designed to encourage work among welfare recipients Man these programs would subsidize individuals only if they work s minimum number of hours. Commonly used techniques cannot direct answers to relevant policy questions since a tied offer is invo and hence the offer cannot be treated as a simple wage change. essence of the problem involves utility comparisons between two or discrete alternatives. Such comparisons inherently require inform about consumer preferences in a way not easily obtained from ord labor-supply functions.

To make such comparisons, I present a method for directly estin consumer indifference surfaces between money income and nonn time. Once these surfaces are determined, they can be used to com variety of alternative programs to investigate whether or not t scope for Pareto-optimal redistribution of income transfers and improving the general level of welfare of the community at large w reducing the welfare of individuals receiving income transfers. Kno of these indifference surfaces allows us to estimate reservation w estimate the value of nonworking-women's time (Gronau 1973), force participation functions, hours-of-work functions, and welfar due to income tax programs (Harberger 1964). I demonstrate tha estimation of indifference surfaces allows us, at least in principle,

This research was sponsored by National Science Foundation and Office of Opportunity grants to the National Bureau of Economic Research. I gratefully edge Ralph Shnelvar's skilled programming assistance. Orley Ashenfelter Betsey, Jacob Mincer, Marc Nerlove, Mel Reder, James Smith, Finis W Robert Willis provided useful comments. I thank Elisabeth Parshley for her as retain full responsibility for all errors in the paper. This paper has not underg staff review and is not an official publication of NBER, NSF, or OEO.

the conventional assumption that the wage rate is independent of hours of work The separation of preferences and constraints allows us to estimate labor-supply functions for individuals affected by welfare systems and progressive income taxation

The methodology presented here is similar in spirit to the pioneering work of Wald (1940). In his neglected paper, Wald suggested that information from different price situations can determine a family of indifference curves up to a second-order approximation In this paper, I depart from Wald's methodology and suggest a plausible estimable specification for the marginal rate of substitution function between goods and leisure, the parameters of which are estimated by maximum-likelihood methods.

The particular focus in this paper is on the effect of work-related child-care programs However, the methodology is more general and may be applied to a variety of work-subsidy programs

I. An Anatomy of Proposals for Child-Care Programs

Proposals designed to relieve work-related child-care expenses have received the most attention The Nixon administration has promoted several measures of this type. In 1971, an administration bill (HR 1) was introduced as "workfare" legislation designed to vacate welfare rolls by providing child care to working women In 1972, tax laws were modified to give generous deductions for work-related child-care expenses if a woman worked 30 hours a week or more. Since some work requirement seems likely in any future legislation, I confine my analysis to such programs

If a child-care program gives a woman with a child an hourly supplement for each hour she works, and she is free to spend it on any child-care source, the supplement is equivalent to a wage change of equal magnitude on both her hours of work and her decision to work. Given reliable estimates of work participation and hours-of-work functions, it is straightforward to make projections of the labor-supply effects of alternative programs.

If a cash grant or tax rebate is given to working women with children, it is no longer possible to proceed so simply Given an agreed minimum number of hours which defines the condition of working, the wage rate for hours worked in excess of the minimum is not affected by such offers although the wage rate for hours worked below the minimum is increased For currently working women who work in excess of the minimum, there are only income effects. If nonmarket time is a normal good, such offers will diminish their hours of work. Women working below the minimum will either increase their work to the minimum number of hours or will be unaffected by the program. For nonworking women market work is

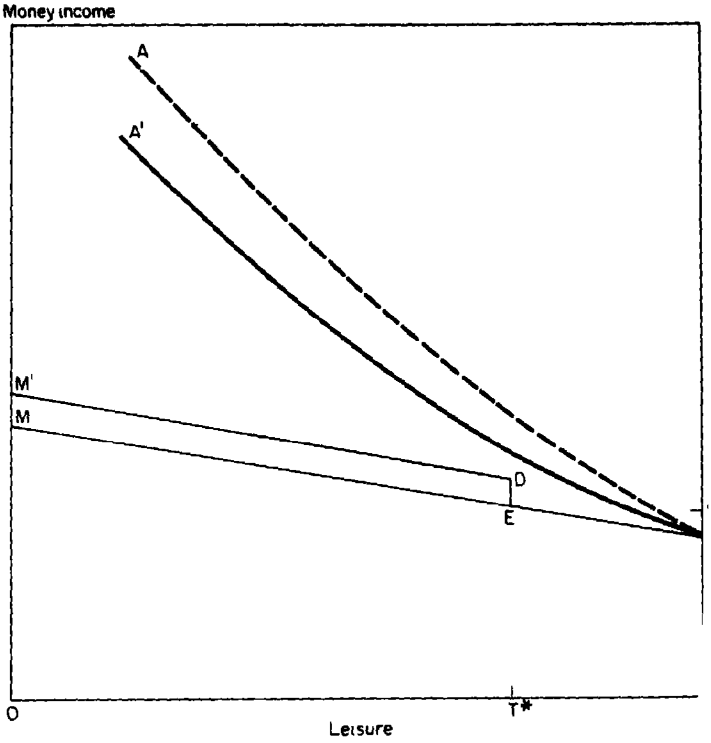


FIG 1

more attractive, and some portion of them will be expected to comm working.

Knowledge of the reservation wage (i e , the minimum wage th woman will accept to go to work) does not enable us to form estima program labor-supply effects To see this, consider the familiar l leisure diagram reproduced in figure 1

Since we only consider wage changes, we may compress all other into money income and construct an indifference curve between m income and nonmarket time or leisure. That there are many us nonmarket time is inessential to the argument as long as time is transferable among its uses.

Suppose that a woman (or family unit) can enjoy *TC* units of consumption if she does not work and that her wage rate is indep of her hours of work so that the market opportunity line is *CM* individual (or family) is on indifference curve *AC* if the woman d work. As drawn, the slope of the indifference curve at zero hours (i.e., the reservation wage) exceeds the offered wage rate (given slope of *MC*) so that the woman does not work. Suppose the gove gives the woman supplement *SC* but requires her to work at least T^*T hours to qualify as working. As the figure is drawn, the woman will not take the offer. Her effective budget constraint becomes *M'DEC*, and the government subsidy does not improve her (or her family's) welfare. Raising the size of the subsidy or reducing the extent of required effort may reverse this conclusion

In practice, tax rebates may be more complicated If the "work requirement" becomes a minimum-earnings requirement which must be attained to be able to deduct child-care expenses, T^* will depend on the wage rate, shifting to the right for higher-wage women and making the program differentially more attractive to them, thus working in an opposite direction to the Nixon "workfare" proposals, *ceteris paribus* Furthermore, since high-income families are more likely to itemize deductions, tax rebates for child care will tend to favor such families, although surely in the presence of these rebates more low-income families would find it profitable to itemize their deductions.

Suppose that the government operates on the supply side of the market by reducing the price per unit quality of work-related child care [1] Remembering that the composite commodity theorem (Hicks 1946) allows us to aggregate all goods whose relative prices do not change into money income, a reduction in the price of work-related child care is equivalent to a shift in the indifference map of figure 1 from *AC* to *A'C* [2] With lower price per unit quality, consumers can attain the same basket of market goods (including the quality of their work-related child care) with lower money income Hence, *A'C* lies below *AC* but the two curves meet at *C* because at this point no work is undertaken and the expenditure on work-related child care is zero Its price is irrelevant to the height of the curve at this point but a reduction in the price of work-related child care reduces the slope of the curve at *C* (i e , the reservation wage) [3]

---

[1] I ignore the practical difficulties in enforcing child-care provision solely for working time

[2] This analysis assumes that individuals are free to choose the quality per unit hour of their child care If for some reason the quality per hour is not an object of choice, the appropriate wage for labor-supply decisions is the wage net of hourly child-care costs Only in this case can child care be treated as a cost of work Since fixity here, as elsewhere, is unreasonable to impose a priori, I do not pursue this approach However, I mention this case because it is tempting to treat child-care expense as a cost Throughout this paper I assume that quality can be varied That only rarely do children go unsupervised when a mother works is a statement about consumer preferences and not one about the need to consume a minimum amount of child-care quality Using these expenditures as exogenous variables in regression analysis of labor supply when in fact they are an object of choice would bias the resulting labor-supply estimates

[3] The proof of these propositions is relatively straightforward and is deleted for the sake of brevity. Note that, in the text, I only argue that the reservation wage for *AC* exceeds that for *A'C*. In general, the slope of *AC* might not exceed the slope of *A'C* for any arbitrary hours-of-work position unless total expenditure on child-care quality rises with increasing hours of work along any iso-utility curve or, what amounts to an equivalent proposition, that work and quality per hour of purchased child care are Hicks-Allen substitutes.

Since the reservation wage has decreased more women will now work. If the price per unit quality is reduced for all child care, whether or no it is work related, the two curves no longer intersect at *C*, but if the expenditure on nonwork-related child care is small the assumption of intersection may remain plausible.

An important question for policy purposes is whether women receivin grants should be free to spend those grants on any form of child car service. In President Nixon's veto message (December 10, 1971) a unrestricted voucher scheme was proposed. However, the Secretary of tl Department of Health, Education, and Welfare testified that tl administration favored monitoring voucher use so that only "authorizec quality sources could be used for expenditure of child-care vouche (March 27, 1972) [4] This distinction is quite important It is well knov that over 80 percent of working women with small children do not u formal day-care services (Low and Spindler 1968; Ruderman 196 Since families are free to resort to the market, the implication of this f is that the price per unit quality of these informal sources is less than tl available in the market If individuals are given cash grants or w supplements and simultaneously are allowed to spend them only in formal market, two offsetting effects are at work for women who h informal child-care sources available On the one hand, a wage or inc subsidy for working raises the attractiveness of work effort On the o hand, the price change shifts the indifference curves (e g , from *A'C* to in figure 1) and tends to offset work incentives A currently worl woman who uses informal arrangements may either remain unaffe by the program or switch into the "formal" program For policy purp it is important in assessing revenue costs to know how many will, in switch over to formal sources

II. Estimating Indifference Curves

If we knew a consumer's system of indifference curves and how system shifts in response to variations in the price of child care, we answer the policy questions raised in Section I Since ordinary l supply functions are derived from indifference curves and their parar may be used to estimate such curves, why bother to directly est them? My answer comes in four parts.

The first is that direct estimates of indifference curves allow us to the parameters of both hours-of-work and decision-to-work fur from a common set of parameters, enabling us to present a unified work in which to interpret these two aspects of work behavior as giving an economy in parameters to be estimated. Second, by esti indifference curves we can combine observations on all women, whether or not they work, and can avoid both the extrapolation error involved in using observations on working women to estimate the indifference surfaces appropriate to all women and the censored-sample problem that arises from estimating labor-supply parameters on a subsample of working women A sample is censored when data are missing for some observations for reasons related to the model in question. Thus, for example, when a woman does not work we do not know what wage she would earn were she to work The reason she does not work is that her reservation wage (i e., the slope of the indifference curve at the zero hours-of-work position) exceeds the (unknown) available market wage Estimates of wage functions and labor-supply functions based on subsamples of working women lead to biased parameter estimates (see Aigner 1971; Heckman 1972; Gronau 1973) Third, by estimating indifference surfaces we can directly estimate reservation wages and can use these estimates as Gronau (1973) has done to estimate the value nonworking women place on their time These estimates provide a "reasonableness" criterion with which to judge labor-supply estimates apart from the usual sign restrictions on income and substitution effects We can further judge the desirability of any labor-supply specification by asking what reservation wage the functional form implies

---

[4] Again, I abstract from the very real problem of enforcing quality standards

A fourth reason for directly estimating indifference curves is that the procedure allows for the separation of preferences from constraints The ability to make this separation is less important if we are willing to make the conventional assumption that wage rates are independent of hours of work (e g , Kosters 1966) but becomes quite important when we acknowledge the existence of progressive taxation, welfare regulations, and time and money costs of work Thus, for example, even if the pretax wage is independent of hours of work, progressive income taxes create breaks in the budget constraint, since tax rates are set on income intervals and change discontinuously at boundaries While it may be reasonable to assume smoothness in tastes, the resulting labor-supply functions may be quite intractable either to specify or to estimate (see Wales 1973) My procedure supplies a natural solution to this problem, and it admits of ready generalization to data from welfare populations which face similarly distorted budget lines.

A. Estimable Indifference Curves

Any indifference curve may be characterized as a locus of points or as an envelope of tangents. For my purposes, the latter characterization is more convenient. Given any initial consumption position, the reservation wage is the slope of the indifference curve at zero hours of work. If leisure is a normal good, higher initial endowments lead to higher reservation wages (Hicks 1946, pp. 28–29). We have seen that higher prices for child-care quality lead to higher reservation wages if a woman uses some child care when she works.

We may write the marginal rate of substitution function (or slope of the indifference curve) at a given level of prework income Y as

$$m = m(p_q, Y, h), \qquad (1)$$

where we ignore variations in the price of other goods and where h is hours of work and p_q is the price per unit quality of child care. A consume possesses a family of indifference curves indexed by level indicator Y, th no-work level of money income or consumption. We know that $\partial m/\partial Y >$ ((normality of nonmarket time) and that $\partial m/\partial p_q > 0$ at the zero hours of-work position. Moreover, it is plausible that this inequality remain valid for all values of h [5] From diminishing marginal rate of substitutio between goods and time we know that $\partial m/\partial h > 0$

If a consumer faces a parametric wage W, at initial income position she works if $W > m(p_q, Y, 0)$. If this inequality applies, the equilibriu position is characterized by

$$W = m(p_q, Y^*, h^*), \qquad ($$

where Y^* is a level index appropriate to the equilibrium indifferer curve and is nothing more than the amount of money income that wo make the consumer indifferent to a choice between working h^* hours wage rate W to gain total resources $Wh^* + Y$ or not working and ceiving income Y^*. As things stand, without knowledge of Y^*, we can deduce the equilibrium relationship between W and h However, gi Y^*, we know that equilibrium also requires that

$$Wh^* + Y = \int_0^{h^*} m(p_q, Y^*, h)dh + Y^*.$$

This equilibrium condition states that we may imagine moving consumer to her final equilibrium position in one of two ways: givin a flat hourly wage rate of W or giving her payment $Y^* - Y$ to be a to her initial resource endowment Y and compensating her by amo

$$\int_0^{h^*} m(p_q, Y^*, h)dh,$$

which is nothing more than the area under her (real) income-compe supply curve for labor. Given W, h, Y, and p_q we may solve out fc

---

[5] Strictly speaking, it is possible that beyond certain values of h this expression b negative, but for this to occur would require decreasing quality expenditure o care as hours of work increase, holding the consumer at the same level of uti n. 3 above).

From equation (2), if m is monotonic in Y^* we may solve for Y^* as a function of p_q, W, and h: $Y^* = g(p_q, W, h)$; using this value in equation (3), we implicitly define the labor-supply function by

$$Wh + Y = \int_0^h m[p_q, g(p_q, W, h), h]dh + g(p_q, W, h)$$

B A Particular Functional Form for m

A wide variety of functional forms may be used to specify m[6] One plausible specification, arrived at after considerable empirical experimentation, is to write

$$\ln m = \alpha_0 + \alpha_1 \ln p_q + \alpha_2 Y + \alpha_3 h + \alpha_4 Z + u \qquad (4)$$

where, as before, p_q is the price per unit quality of child care, Y is the prework level of income, h is hours of work, and Z is a vector of constraints to be discussed more fully below A random variable designated "u," with zero mean and variance σ_u^2, reflects variation in preferences for work among individuals The previous analysis leads me to the prediction that $\alpha_2 > 0$ (normality of leisure) and $\alpha_3 > 0$ (diminishing marginal rate of substitution between goods and leisure) and to the presumption that $\alpha_1 > 0$

Manipulation along the lines discussed in section A shows that the resulting labor-supply function is implicitly defined by

$$\begin{aligned} u = {} & \ln W - \alpha_0 - \alpha_1 P_q - \alpha_3 h - \alpha_4 Z \\ & - \alpha_2 \left[Wh + Y - \frac{W}{\alpha_3}(1 - e^{-\alpha_3 h}) \right] \end{aligned} \qquad (5)$$

This labor-supply curve can become backward bending beyond a certain value for hours worked[7]

This particular functional form is offered as a starting point If it is grossly inconsistent with the data, either α_2 or α_3 will be negative, and the implied reservation wages, predicted from estimates of equation (4), will be unreasonable

---

[6] One method is to let the data determine the functional form, in the manner of Box and Tidewell (1962) and Box and Cox (1964) I did not pursue that approach in this paper

[7] Straightforward differentiation of eq (5) shows that $\partial h/\partial W = \{(1/W) - \alpha_2[h - (1/\alpha_3)(1 - e^{-\alpha_3 h})]\}/[\alpha_2 W(1 - e^{-\alpha_3 h}) + \alpha_3]$ Clearly, $1/W$ is nonnegative, the denominator is positive, as is the first expression in brackets To see why the latter assertion is true, note that when $h = 0$ the first bracketed expression is zero and that the partial of that bracketed expression with respect to h is $(1 - e^{-\alpha_3 h})$, which is clearly nonnegative since α_3 is positive and $h > 0$ Thus, it is possible, for suitably large values of h, for $\partial h/\partial W$ to become negative.

C. *The Empirical Specification*

To estimate the relevant parameters, we must specify the time dimension of the decision period and the vector of constraints (Z) which affect the marginal rate of substitution between goods and leisure. For the purposes of this paper, I use hours worked in a given year. In my judgment, this dimension is the appropriate one for analyzing the labor-supply behavior of women with children—the target population for child-care programs—since we are interested in knowing how they will respond to programs at the age when they have their children. I simplify the analysis by assuming that variations in the wife's labor supply do not affect the husband's earnings.[8]

We may follow a well-established tradition in the literature and postulate a one-period static model, or we may imagine the individual woman (or household) maximizing her lifetime utility function at each age of her life, subject to the constraints reflecting past decisions and chance events. The traditional static model may easily be shown to be a special case of a more general life-cycle model.

To simplify exposition and maintain comparability with previou studies in labor supply, I assume a one-period framework After presentin the essential features of my procedure under this assumption, I then shov how to incorporate more general intertemporal considerations.

In specifying the list of variables (Z) that would plausibly affect th marginal rate of substitution between goods and leisure, the presence children of different ages and the price per quality unit of child care a the most obvious candidates

Since children require time and effort, it is likely that the presence children, especially young children, makes market work less attractive If a mother works and values her child's environment during her worki hours, she will make child-care arrangements. The effect of children the indifference system allows us to estimate what compensation necessary to induce the woman to work and how the presence of child affects the hours worked by women.

An important empirical fact is that most working women with sn children do not pay for their child care, or if they do pay they use formal sources and pay less than women using formal sources. S economists have interpreted this fact as evidence of failure in the c care market.[10] I interpret these facts differently. It is plausible that some women have low-cost, nonmarket substitutes available to care for their children. Potential sources of low-cost child care are relatives living nearby, husbands with flexible working hours, and friends and neighbors The availability of these low-cost sources lowers the reservation wage and makes it more likely that women having such sources available will work.

---

[8] This assumption is widely used (see Gronau 1973; Bowen and Finegan 1969), is clearly open to the criticism that the family labor supply is jointly determined. assumption is equivalent to fixing the husband's nonmarket time institutionally cluding it altogether from household preferences.

[9] For example, if children require at least some minimum amount of the mother's then presence of a child, viewed solely as a reduction in available time without setting utility flow, would raise the reservation wage as long as consumption is a good.

Unfortunately, it is not possible to directly measure the price per unit quality women must pay for their child care It is plausible to argue that all families can resort to the market to purchase units of child-care quality but that some families have informal sources available at a lower price per unit quality than is available in the market Measuring quality by expenditure, we may normalize the market price at unity [11]

Suppose the distribution of informal prices among the population at large is given by

$$\ln p_q = X\beta + \varepsilon \tag{6}$$

where ε is a disturbance with zero mean and variance σ_ε^2 and where X is a vector of variables assumed to affect the price per unit quality of child care. If we assume that the family selects its source of child care by the rule min (0, $\ln p_q$), the marginal rate of substitution function at the prework level of utility for families using informal sources ($\ln m_I$) may be written as

$$\ln m_I = \alpha_0 + \alpha_1 X\beta + \alpha_2 Y + \alpha_3 h + \alpha_4 Z + u + \alpha_1 \varepsilon, \tag{7}$$

while the function for women using formal sources ($\ln m_{II}$) may be written as

$$\ln m_{II} = \alpha_0 + \alpha_2 Y + \alpha_3 h + \alpha_4 Z + u \tag{8}$$

where a restriction on the distribution of ε arises from the condition that

$$\ln p_q < 0 \rightarrow X\beta + \varepsilon < 0 \tag{9}$$

It is possible to estimate β/σ_ε from data on whether or not a family uses informal child care when the wife works However, further information is required to estimate the coefficient α_1 alone,[12] although it is possible to estimate $\alpha_1\sigma_\varepsilon$ Since σ_ε is known to be positive, it is still possible to test the hypothesis that α_1 is positive

[10] In particular, Nelson and Krashinsky (1972) argue this position

[11] This procedure is clearly open to debate but is conventional It assumes that quality is an objective characteristic and that units of quality are available at prices independent of the level of quality or the number of children.

[12] To see this, note that we can estimate $\alpha_1\beta$ from eq (7) and $\alpha_1^2\sigma_\varepsilon^2$ Suppose, as is asserted in the text, we know β/σ_ε. Then, from the "intercept term" we can estimate $\alpha_1\sigma_\varepsilon$, and from the variance we can also estimate $\alpha_1\sigma_\varepsilon$ if ε is independently distributed from u, but we cannot determine α_1 independently from σ_ε unless we assume further information One possible assumption is that $\sigma_\varepsilon = 1$

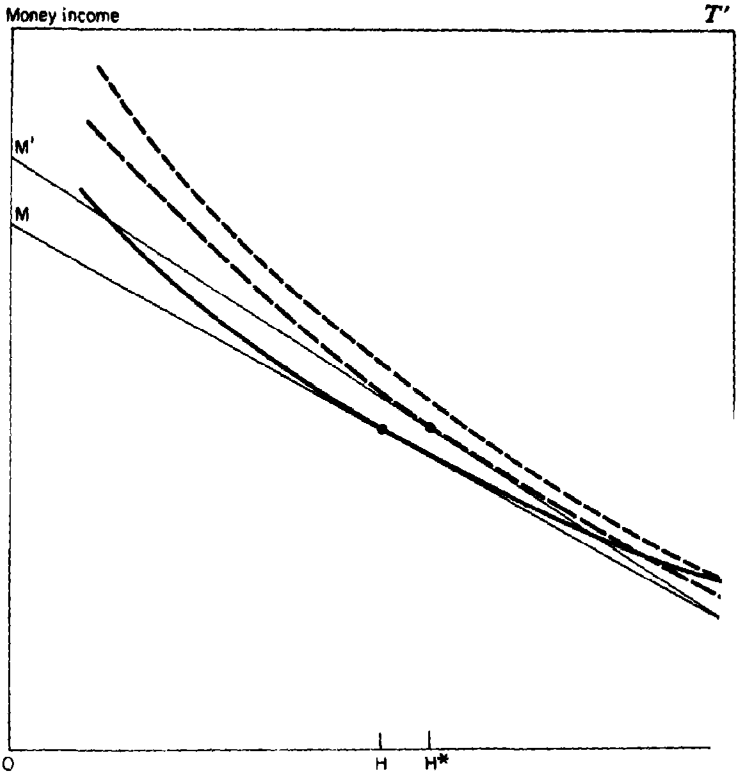


FIG 2

Suppose that we can successfully estimate the relevant paramete equations (7) and (8), including $\alpha_1\beta$ We can use these functio address the problem, raised in Section I, of restricted vouchers. government program raises the wage rate of a woman with a chil insists that she use formal child-care sources, it is equivalent to a wage change for a woman already using formal sources but invo change in p_q to unity (ln p_q to zero) for a woman not using these so From estimates of the indifference surfaces we can determine whet not (on average) a woman using informal sources would switch to sources, because we can estimate the equilibrium position for h work in the case where she gets a higher wage but must pay a highe for child care and the case where she abstains from the program. determined her equilibrium hours of work and wage rate in each sit we can use equations (7) and (8) to determine the Y^* (the am money income without work that would put the women at the sar of utility as working would) implicit in each situation. Since Y^* serves as an index of utility, the situation with the higher Y^* is the preferred situation and will be chosen This approach is exact only if a woman spends nothing on child care when she does not work; it is approximately correct if child care is a small fraction of family expenditure when the woman does not work

To see this more explicitly, consider figure 2 which, is similar to figure 1 Consider a woman facing budget constraint MB and working TH hours. The government raises her wage but simultaneously increases the price of work-related child care. In terms of this diagram, the effect of the price change is to shift "preferences" as shown by the dashed lines and to raise the wage so that the new budget line becomes $M'B$. A new equilibrium position emerges along isoquant Y^*, using the "new" system of preferences Using the estimated labor-supply curves, we can deduce that TH^* hours would be worked Further, using the new equilibrium values, we can deduce level Y^* from equation (8) Then we can determine which situation is preferable since it is assumed that child-care expenditure is a negligible portion of consumer expenditure if a woman does not work, so that the effect of changes in the price of work-related child care is to pivot the indifference map around the intercept on the TT' axis As the diagram is drawn, the woman abstains from the program since Y^* lies below Y' Precisely the same methodology may be used in analyzing tied offers

D Life-Cycle Modifications

Assume an additively separable lifetime-preference function which may be represented as the (time preference) discounted sum of utility functions at each age [13] In this specification, utility at one age is not directly affected by variation in the consumption of time and goods at other ages However, given a lifetime budget constraint, increasing consumption at one age reduces resources available at all other ages

Within the context of this model, it is heuristically convenient (and formally correct) to imagine the consumer maximization process as a two-stage affair: subject to any income allotment, the family maximizes its utility within each period To determine its allotment among periods, the family allocates income to equalize the marginal constribution to lifetime utility of an additional dollar of income in all periods if it is free to transfer resources among periods at a given borrowing and lending rate.

To account for the mobility of funds among periods, we must allow for saving or dissaving in a given period. For example, by borrowing against future income, households may enhance current resources. By saving, households transfer some of the current flow of resources to future time periods. Accordingly, the budget line in the conventional labor-leisure diagram (see fig. 1) must be shifted up (down) by the amount of net debt accumulation (saving) in the period. If saving is set at zero, the traditional one-period model retains its validity in a life-cycle context.[14]

---

[13] This specification is widely used in the optimal-growth literature (e g , Shell 1967) and the operational-utility literature (e g , Strotz 1957)

Assuming that, there is no saving or borrowing is restrictive but traditional. It simplifies the analysis but is not necessary to it. Since the value of saving at any age depends on all prices, wages, and exogenous incomes, including the current wage, the current wage enters the analysis in two ways: as a direct determinant of work effort through the slope of the budget constraint, as in the traditional analysis, and as a determinant of the level of the given period's budget constraint through its effect on saving. Thus, in addition to the usual income and substitution effects confined to one period, there is an intertemporal income-transfer effect of wage change.

Introducing intertemporal considerations alters the simplicity of some of the previous analysis. In particular, consider the effect of a government program which raises the wage rate for women at an age at which children are present and simultaneously requires that the woman use approved child-care sources In a one-period context we are able to solve the implicit level of utility, indexed by Y^*, under each program and are able to conclude whether or not a woman participates in such a program. In an intertemporal context we lose this simple feature because it is possible for Y^* to be lower in one time period but for the resulting earnings flow to be sufficiently higher to more than compensate the woman by enhancing her future utility. Thus, we encounter an intertemporal index-number problem.

To illustrate this point, consider figure 3 The dotted indifference curve represents the goods-leisure trade-off at a given period The solid curves are the indifference system after the price per unit quality is raised Consider a woman initially in equilibrium at point B', where TK is the flow of earnings from other sources minus saving in the period. Holding saving fixed, raise the wage so that the new equilibrium is B''. As drawn, B'' is at a lower level of current utility than B'. Yet, if earnings are increased sufficiently, the woman might still opt for the program, since she can transfer resources to future time periods and enjoy greater utility. If this is the case, the budget line shifts down by the amount of saving

---

[14] For a proof of these propositions, see Heckman (1971) where the Strotz "utility tree" is applied to an intertemporal labor-supply problem The assumption of additive separability yields a traditional labor-leisure diagram for each period in the consumer's life cycle. The level of the budget constraint in each period is determined, in part, by exogenous income flows in that period and the net transfer of resources into or out of the period.

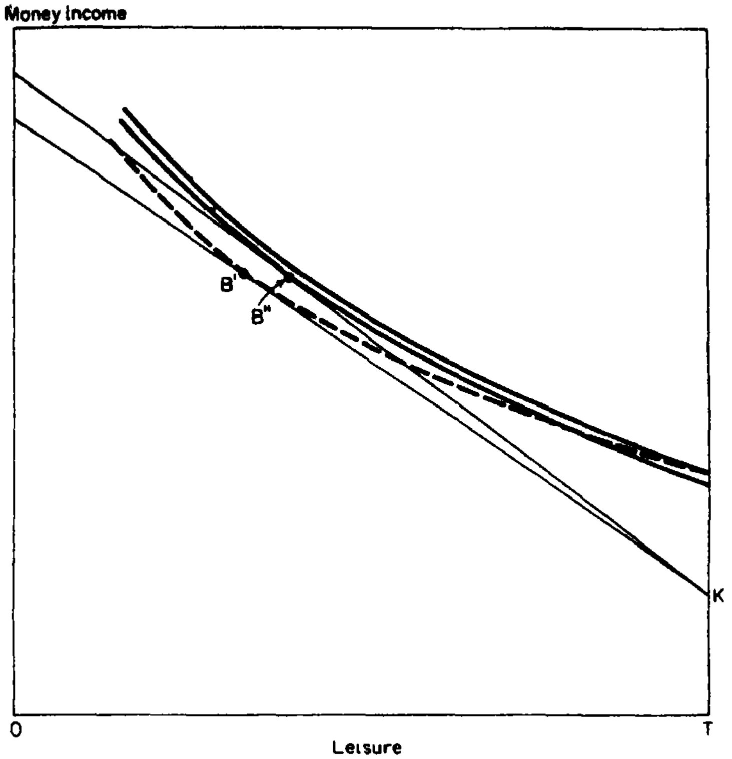


FIG 3

and the woman works more since leisure is a normal good in each time period. Note, however, that if saving is held fixed, and the wage increase leads to a higher level of utility in this period, the possibility of transferring funds among periods enhances the attractiveness of the program so that a woman will definitely choose to participate in it and thus we can estimate a lower bound on the number of program participants. Note, too, that we can estimate an upper bound: take all women who are made worse off in this period and assume they participate in the program. Solving their new equilibrium values, allowing for a transfer of resources among periods, we can solve for their new earnings. If we also estimate their expenditure on child care, their earnings net of child-care expenditure (in the new program) must increase for resources to be freed to transfer to subsequent periods Then we know that only women who have such options will participate in the program, although clearly not all of these women will participate.

The only necessary modification in the previous empirical specification is the introduction of a savings function which is subtracted from the exogenous income flow Y. In the absence of data on saving, we are forced to estimate an implicit savings function.

III. Estimates

Estimates of the one-period model developed in Section II are presented below. The estimation of the complete life-cycle model is left for another occasion [15]

Before presenting the estimates, it is useful to stress the restrictive nature of some of our assumptions. The assumption that the wage rate is independent of hours of work ignores the effect of progressive taxation However, as noted in Appendix B, this defect can be remedied. A potentially more serious problem is our treatment of employer behavior. If firms incur fixed costs per worker, they are not indifferent to alternative combinations of men and hours per man which yield the same total man-hours (Rosen 1968) Under these circumstances, firms might pay higher hourly wage rates for longer weekly hours, at least up to certain levels of weekly hours worked Thus, it is possible that the equilibrium condition invoked in Section II, that wage rates equal the marginal rate of substitution for working women, might not apply This consideration creates a potentially important complication which is ignored in this paper [16]

As Gronau notes, both the presence of fixed costs of work and the possibility of endogeneity of husband's earnings create potentially important biases (Gronau 1973, p S177) These factors could be accommodated within our framework but only at much higher computational cost.

The data used to form the estimates are from the National Longitudinal Survey (NLS) of 1966 for women aged 30–44 The functions are estimated on a subsample of women who were married, spouse present, with at least one child under age 10. This subsample is chosen for two reasons The principal reason is that it is a population obviously affected by child-care programs.[17] Previous studies (Ruderman 1968; Bowen and Finegan 1969; Gronau 1973) suggest that only younger children exert an important retarding effect on women's work effort A second reason for this choice of sample is closely related to the first. The NLS data give information on the availability of informal sources of child care, defined as care by relatives or zero (dollar)-cost day care. A unique feature of these data is that, in addition to questions asked on the type of child care used by working women, questions concerning the availability of informal sources were asked of nonworking women as well It is computationally and conceptually more difficult to incorporate in our sample women for whom child care is an irrelevant issue

---

[15] An attempt was made to estimate a full-fledged life-cycle model on a cross section of data. In principle, this is a simple exercise In practice, since the necessary savings data are missing, it is necessary to specify a savings function, which depends on current and future wage rates and exogenous income flows and on current-asset variables The additional wage and income variable introduced in eq (5) as determinant of saving led to unstable estimates In fact, plausible alternative specifications for the savings function lead to underidentification for some parameters These results suggest that we are asking too much of the data from one cross section. When the second-year data from the National Longitudinal Survey are available, it will be possible to estimate a separate savings function and get reliable estimates of the complete model.

[16] If there were a fixed standard work week, the analysis of Section II could be generalized to accommodate this complication. For those working fixed hours, and not moonlighting, the marginal rate of substitution evaluated at the standard work week (which depends, among other things, on the income flow from work and length of time at work) is less than the potential moonlighting wage, while for moonlighters the opposite is true. Thus, as shown in Appendix B, it is possible to gain some information about consumer preferences even when the marginal equality conditions no longer apply.

In Appendix A I present an estimation method that incorporates the information that (1) the reservation wage is less than the offered wage and (2) the price per quality unit for child care is less than that available in the market Both types of sample-inclusion criteria create the potential of censoring bias, but this is avoided by the procedure used

By the construction of the sample, the presence of at least one child younger than age 10 is assured To allow for scale effects, I created separate variables for number of children aged 0–3, 4–6, and 7–10 Other children variables led to unstable estimates Husband's earnings are added to net worth multiplied by an estimable flow coefficient to approximate the flow of exogenous income relevant to a one-period model

Michael (1973) has argued that education may affect the household's ability to organize and produce its final consumption To accommodate this hypothesis, I present estimates which measure the effect of the education (in number of years) of the wife on the marginal rate of substitution between goods and leisure I contrast these results with more conventional estimates which exclude education from household preferences

In order to include in the sample nonworking women for whom wage data are missing, it is also necessary to estimate a function for the market wages of all women. Its specification is kept simple The natural logarithm of wages is postulated to depend on education and labor-market experience [18] I explicitly allow for correlation between the disturbances of the preference function and the wage function.

A novelty of the procedure used is the explicit estimation of an informal child-care price function (discussed in Section II) Previous studies (e g.,

---

[17] However, the most relevant policy population is women on welfare who have young children. Small sample size and ignorance of the features of the welfare system confronting poor mothers precluded estimation on this sample

[18] For a discussion of the experience variable, see the paper by Mincer and Polachek in this volume. I aggregated their segments of experience into a total-experience variable.

Bowen and Finegan 1969) have shown that children aged 14–18 serve as stimulants to married female labor-force participation, and it is plausible that they supply a ready source of child care. Similarly, it is plausible that relatives living in the home, husbands with flexible working hours, and friends serve as low-cost sources of child care (see Low and Spindler 1968; Ruderman 1968) Accordingly, these determinants of the quality-adjusted price of day care are incorporated in the estimation.

To capture the effect of relatives on the price of child care, I created a dummy variable which assumes the value of one when a sister, parent, or grandparent lives in the household The number of children aged 14–18 is included as a variable. As a crude measure of the flexibility of the husband's hours and his availability for child care, I use the number of hours he works in a week. As a measure of the availability of low-cost care from friends and relatives living nearby, I use two variables a dummy which assumes the value of one when a woman has lived in a Standard Metropolitan Statistical Area (SMSA) all her life and a continuous variable measuring the number of years a woman has lived in an SMSA Both are expected to depress the price of child care.

The estimates are presented in table 1 Column 1 reports estimates of the parameters for whites obtained by excluding the effect of education from household preferences, while column 2 shows the effect of including education Comparable results for blacks are in columns 3 and 4.

Consideration of column 1 coefficients for whites will illustrate the meaning of the coefficients. A unit increase in the logarithm of the quality-adjusted price of child care divided by its standard error raises the marginal rate of substitution by 14 3 percent Each $10,000 of exogenous income raises the marginal rate of substitution by 22 percent Each 1,000 hours worked raises the marginal rate of substitution by 51 percent. An additional child aged 0–3 raises the marginal rate of substitution by 24 3 percent; an additional 4–6-year-old child raises the marginal rate of substitution by 8 4 percent, while an additional child aged 7–10 raises it only 2 percent The coefficient on assets is a number converting the stock of net worth into a flow yield. Thus, if a family has $10,000 worth of assets, the estimated income flow is $380 This flow is added to other exogenous income flows in calculating the effect of income on the marginal rate of substitution.

Similarly, the coefficients for the quality-adjusted price of child care show the effect of the associated variables on percentage changes in this price divided by the standard error of the price distribution. Older children, relatives living in the home, and length of residence in the SMSA all have the expected negative effects on this price while the other coefficients cannot be deemed significant by conventional standards.

The estimates for the market-wage function show the effect of the associated variables on the natural logarithm of market wages. The population correlation coefficient measures the interequation residual correlation between the market-wage function and unmeasured tastes for work.

In comparing columns 2 and 1, it is seen that while conventional significance tests suggest that education should be included in household preferences, its inclusion alters few coefficient estimates. The only notable exception is that the effect of education on market wages is increased and is not significantly different from the effect of education on reservation wages. This suggests that for white women with at least one child younger than age 10, better-educated women do not work in the labor force as much as less educated women. For comparable black women, better-educated women are more likely to work

The most useful way to summarize these numbers is to consider their potential application. Making the bold assumption that the preference parameters for married black women with spouse present are those appropriate to all black women, we can investigate the possibility of using tied child-care offers to alleviate welfare dependency Consider a group of black women currently eligible for an annual welfare payment of $3,000. For the sake of illustration, we assume that each woman has three children—one in each of our estimated age intervals We assume that each woman uses informal child-care sources and that she has one child from 14 to 18 years old, one older relative living in the home, and that she has lived in her current SMSA for 8 years We further assume that each woman has no asset income and 10 years of schooling

Each woman has a choice of working or taking welfare The geometric mean indifference curve appropriate to potential welfare recipients is sketched as AA' in figure 4 Fifty percent of the potential welfare recipients have an indifference curve emanating from A (=$3,000) which lies below AA' Twenty-five percent of the group have an indifference curve starting at A which lies below AC. This latter group may be termed the most "work prone" because they require less compensation to be induced to work Again, solely for the sake of argument, suppose that each woman has an identical wage of $1.50 per hour which she can earn in the market More realistically, we might assume a distribution for wages as well

In this extreme example not all women will choose to go on welfare The few who are not on welfare will be observed to work long hours at low wages. Again, solely for the sake of argument, suppose no women work. Suppose the government offers an unrestricted "day-care" voucher to a woman if she works 1,500 hours From budget line OB we can see that if the woman were to work 1,500 hours at $1 50 per hour she would earn $2,250. The minimum compensation beyond her own earnings that would be necessary to induce half the women to leave welfare is the distance $B''B'$ (=$3,300). Note, however, that if a $2,560 child-care subsidy were offered (=$B'B'''$) 25 percent of the women would leave

TABLE 1

PARAMETER ESTIMATES FOR BLACKS AND WHITES
(MARRIED, SPOUSE PRESENT, WITH AT LEAST ONE CHILD LESS THAN 10)

	WHITES				BLACKS			
	(1)		(2)		(3)		(4)	
Parameters of the marginal rate of substitution								
Intercept, α_0	−0.046	(−2.83)	−1.2453	(3.37)	0.086	(1.19)	−0.233	(−1.56)
Price effect of child care, α_1	0.143	(2.66)	0.110	(2.05)	0.297	(8.11)	0.237	(9.6)
Income term, α_2	0.22×10^{-4}	(2.62)	0.198×10^{-4}	(2.43)	0.13×10^{-4}	(1.43)	0.12×10^{-4}	(14.9)
Hours effect, α_3	0.51×10^{-3}	(4.88)	0.487×10^{-3}	(3.83)	0.43×10^{-3}	(4.2)	0.314×10^{-3}	(5.98)
No. of children aged 0–3, α_{41}	0.243	(3.32)	0.234	(2.95)	0.124	(5.26)	0.076	(5.84)
No. of children aged 4–6, α_{42}	0.084	(2.11)	0.081	(2.59)	0.040	(1.82)	0.022	(1.37)
No. of children aged 7–10, α_{43}	0.020	(0.66)	0.032	(2.09)	−0.036	(−2.1)	−0.031	(1.37)
Wife's education, α_5			0.100	(14.5)			0.054	(2.14)
Flow coefficients on assets, S	0.038	(1.29)	0.024	(0.88)	0.159	(0.05)	0.129	(0.89)
Standard deviation, σ_u	0.668	(6.73)	0.598	(5.1)	0.607	(8.4)	0.492	(14.6)
Parameters of the quality-adjusted price of child care								
Intercept, β_0/σ_ϵ	0.063	(3.43)	0.064	(3.47)	0.023	(0.845)	0.026	(1.00)
Lived in SMSA all your life? β_1/σ_ϵ*	0.069	(0.4)	0.063	(0.365)	0.2459	(1.2)	−0.082	(−0.41)
No. of children aged 14–18, β_2/σ_ϵ	−0.283	(−4.71)	−0.279	(−4.63)	−0.084	(−1.4)	−0.106	(−1.8)
Relative living in home?, β_3/σ_ϵ†	−0.666	(−2.92)	−0.666	(−2.92)	−0.424	(−2.1)	−0.405	(−2.1)
Husband's hours of work last week, β_4/σ_ϵ	0.0012	(0.349)	0.0012	(0.334)	0.003	(0.5)	0.0001	(0.298)
Length of residence in this SMSA, β_5/σ_ϵ	−0.0122	(−2.12)	−0.0120	(−2.09)	−0.0138	(−1.69)	-0.45×10^{-2}	(−0.56)
Parameters of the wage function								
Intercept, m	−0.266	(1.49)	−0.815	(−7.99)	−0.149	(−0.99)	−0.329	(−2.22)
Education, n_1	0.059	(4.45)	0.103	(13.0)	0.0454	(3.37)	0.080	(5.6)
Market experience, n_2	0.026	(8.63)	0.027	(10.3)	0.0137	(7.2)	0.011	(8.53)
Standard deviation, σ_v	0.393	(26.01)	0.381	(28.2)	0.487	(23.2)	0.433	(33.7)
Population correlation between the disturbances of tastes (u) and wages (v), ρ	0.598	(5.36)	0.566	(3.51)	0.85	(20.1)	0.91	(70.4)

TABLE 1 (*Continued*)

	Number of Observations	
	Whites	Blacks
Do not work, would use formal sources‡	248	87
Do not work, would use informal sources	148	62
Work and use formal sources	104	45
Work and do not use formal sources .	293	187
Total	793	381

Sources —Data are from 1967 National Longitudinal Survey of Work Experience of Women aged 30–44 *For a description of these data, see Dual Careers* (Parnes, Shea, Spitz, and Zeller 1970)

Note —Numbers in parentheses are asymptotic normal statistics estimated from the information matrix

* = 1 if yes, 0 if otherwise

† = 1 if sister, parent, grandparent, 0 if otherwise

‡ Formal sources are defined as paid care at child-care centers and care at a nonrelative's home Informal sources include all other sources Nonworking women are classified into two groups those looking for work (group B) and those not looking for work (group C) Everyone in the first group was asked whether or not child-care arrangements would be necessary and what type of arrangements would be made if they were necessary This allowed for a classification broadly consistent with the classification of actual use for working women Information on the group not looking for work is more scarce To be included in the sample they had to answer "yes" or "it depends" to a hypothetical question about whether they would work if a job were offered to them (Parnes et al 1970, question 30a, p 260) and a subsequent question on need for child care and what form would be used (question 31) To determine whether the composition of the sample was biased by using this inclusion criterion, separate discriminant functions were fitted by race for all married women *with at least one child younger than 10 years old in group C Respondents* were *distinguished* from nonrespondents by a set of variables including the number of children in different age intervals, assets, age and education of wife, and husband's earnings The only significant discriminants for both races were assets and the number of children aged 0–3 years. Both variables tended to be larger for nonrespondents. Accordingly, the coefficient estimated reported in this table may be biased In my opinion the added information about potential child-care sources for nonworking women was worth the risk of bias

welfare voluntarily, and the welfare cost to society at large would diminish by $440 per person who leaves welfare [19] Clearly, in this case there is scope for using tied offers to reduce welfare dependency The transfer-minimizing policy is Pareto optimal and is straightforward to determine [20] Note, too, that we can determine what proportion of women will work beyond the minimum number of hours needed to qualify for the program Consider a woman whose indifference curve is AC At 1,500 hours of work, the slope of the indifference curve (at B''') exceeds $1 50, so that if she were given slightly more than $2,540 she would leave welfare and work

[19] Remember that it was assumed that all women were on welfare In reality, only some fraction of this 25 percent would be.

[20] For each value for hours worked, h, a child-care bonus B determines a proportion of welfare clients $P(B)$ who will leave welfare, given that bonus Clearly, B lies between zero and $3,000, the welfare payment The mean transfer to the group is $P(B)B + [1 - P(B)](\$3{,}000)$ Minimizing mean transfer is equivalent to minimizing group transfer since I assume no discrimination within the group is possible For each h, there will exist a B which minimizes mean transfer, and the minimum mean transfer for all h can be determined. Thus, knowing the distribution of preferences, it is possible to simultaneously determine B and h Under general conditions, there exists a unique Pareto-optimal policy for tied welfare offers which can be estimated from the data. Note that, in the example, it is assumed that no one works initially To determine the Pareto-optimal tied offer if some work, it is necessary to subtract the transfers paid to those who would have worked without a child-care bonus. Note that Pareto optimality is used here in the restricted sense of improving community welfare from a current position, it is not the global optimality concept of general equilibrium theory

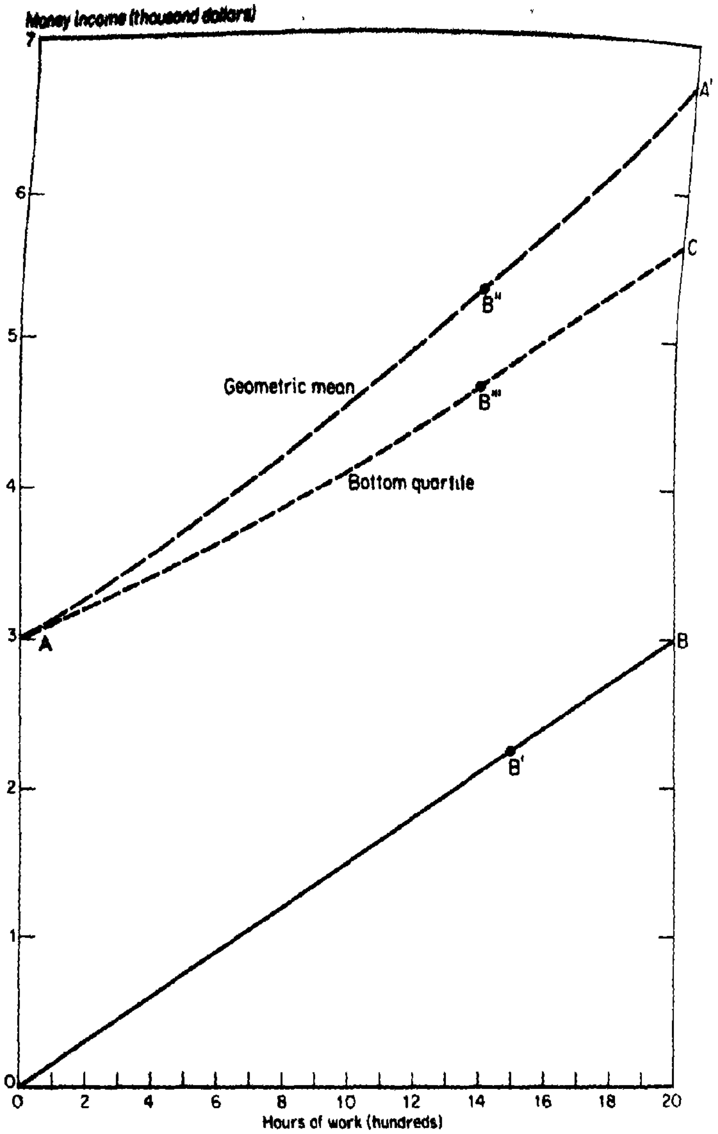

Money income (thousand dollars)
7
6
5
4
3
2
1
0
A'
C
B''
B'''
Geometric mean
Bottom quartile
A
B
B'
0
2
4
6
8
10
12
14
16
18
20
Hours of work (hundreds)

FIG. 4

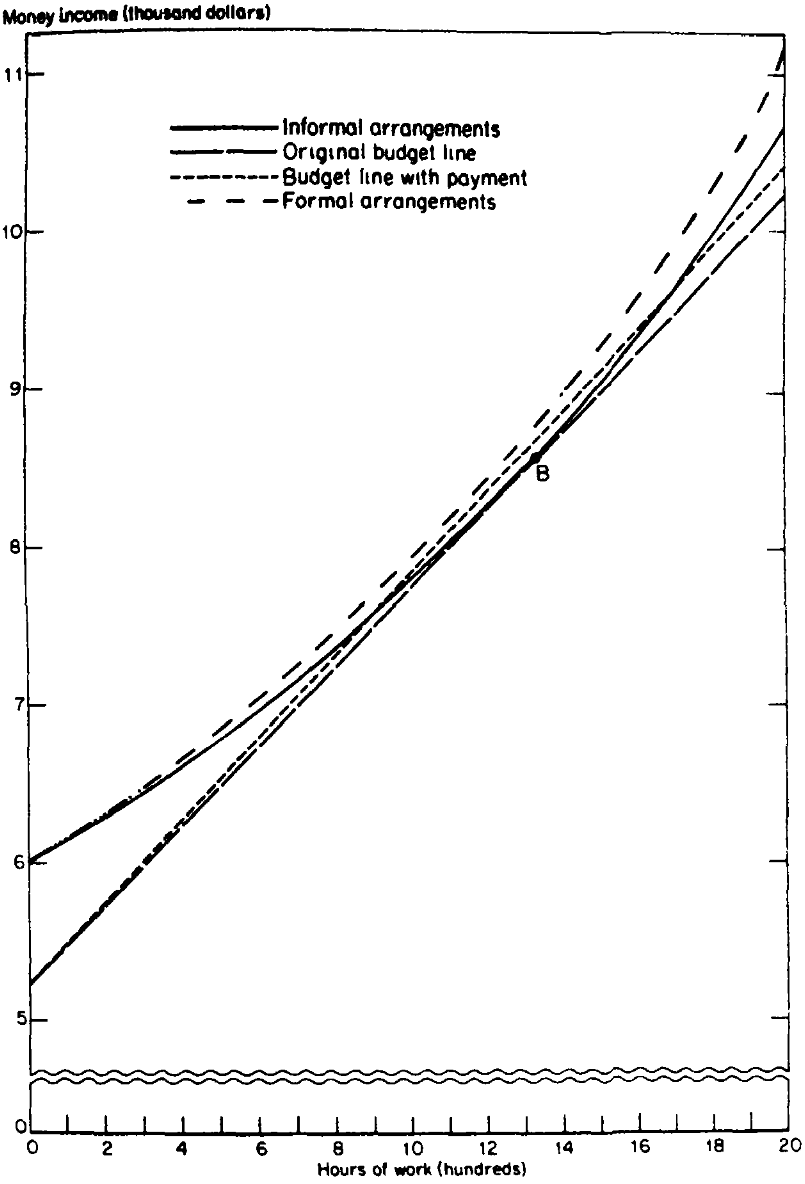

Money income (thousand dollars)
11
10
9
8
7
6
5
0
Informal arrangements
Original budget line
Budget line with payment
Formal arrangements
B
0
2
4
6
8
10
12
14
16
18
20
Hours of work (hundreds)

Fig 5

exactly 1,500 hours. The reason she would not work longer hours is that her reservation wage at 1,500 hours exceeds her offered wage of $1 50

As a further illustration of the use of the empirical results, consider a representative working white woman, using informal child-care sources, whose husband earns $5,500 per annum. Suppose that she has 12 years of schooling, assets of $5,000, and two children, one from 0 to 3 years old and one from 4 to 6. In figure 5, we plot a solid system of indifference curves appropriate to the case of lower prices for informal child care, drawn on the assumption that the woman has not lived in the SMSA all her life, that her husband works 40 hours a week, that a relative lives in the home, and that she has lived in her current SMSA for the last 10 years Superimposed on this system is a broken-line indifference curve appropriate to the use of child care purchased in the formal market

As the lines are drawn, the woman is in equilibrium at point *B* earning $2 30 per hour Suppose the government raises her wage to $2 40 per hour by providing a 10¢ per hour child-care subsidy conditional on her use of formal sources Will she take the offer? In this case, the answer is no The dashed "program budget line" lies below the iso-utility contour for formal arrangements, equivalent in utility terms to the indifference curve for informal arrangements on which she previously was in equilibrium Hence, she abstains from the program Given the parameters of the distributions of the curves, we can make similar statements for groups as a whole and hence can estimate the proportion of working women currently using informal child-care sources who will switch to formal sources if only restricted vouchers are available [21]

IV. Summary and Suggested Extensions

The economics of tied work payments and methods for estimating the effect of such payments on labor supply have been discussed in this paper It is important to distinguish the conceptually easier problem of modeling the response to tied offers from the more demanding problem of providing reliable estimates of the appropriate behavioral functions It has been shown that knowledge of consumer preferences is necessary to estimate program effects, and methods have been suggested for determining these preferences.

By directly estimating indifference curves, hours of work and work-participation equations have been derived from a common set of parameters. The separation of preferences from constraints allows us to estimate the labor-supply parameters of individuals from data generated by non-standard constraints, such as the broken-line budget constraints resulting from the tax system, where a tractable labor-supply function does not exist. At the cost of estimating a savings function, we can embed the traditional one-period model of labor supply into a life-cycle model Both the distribution of tastes for work and the distribution of market wage rates for the population at large are estimated The estimates suggest that wage rates are strongly correlated with preferences for work so that simple "reduced-form" labor-supply functions obtained by regressing hours worked on wage rates give biased estimates In forming estimates, a statistical procedure is employed which avoids this bias and the censoring bias discussed above.

---

[21] To answer both policy questions fully would require consideration of the distributions of wage rates, preferences, and prices for child care Such projections would be suspect at this stage of this research because the full tax-adjusted life-cycle model has not been estimated.

Perhaps the most controversial aspect of this study is the treatment of the choice of the mode of child care A latent distribution of informal prices for child care is assumed to exist in the population at large It is further assumed that a competitive market in child-care quality exists so that individuals facing informal prices in excess of the competitive price use formal child-care sources if they work The decision to work and the decision to use formal sources are related in the sense that the price of child care is a determinant of the decision to work and of the actual hours worked This model enables us to use information on hours of work and the decision to work, in combination with direct information on the form of child care used by working mothers, to estimate the determinants of the mode of child care mothers select The empirical results are consistent with this model but have not been tested against possible alternative models

If families value the development of their children, and the mother's time is an important input into this development, custodial child care might be chosen if a mother works a limited number of hours but might be unacceptable if she works full time. If formal sources provide higher-quality day care, the decision to work, the length of the work week, and the choice of the mode of child care are jointly determined and mutually dependent on the wage rate and the prices of formal and informal care Although this alternative model is much more difficult to estimate, it is conceptually more attractive and contains the present model as a special case For these reasons it is an approach that might prove fruitful in future research on the economics of child care

Appendix A

The Estimation Procedure

The parameters of the marginal rate-of-substitution function (4) are estimated by a maximum-likelihood technique developed in a previous paper (Heckman 1972). In that paper, I present a method that permits observations on nonworking women to be pooled with observations on working women to estimate hours-of-work functions, work-participation functions, and wage functions for all women. The procedure avoids the bias that results from estimating wage and hours-of-work functions using standard regression techniques on censored subsamples of working women.

Four distinct subsamples are considered here in contrast to the two subsamples considered in my previous paper. Some working mothers use informal child-care sources while others use formal sources. Potentially, the same dichotomy exists among nonworking women, and a unique feature of the data used here is that for many nonworking women it is possible to identify which mode of child care they would use if they were to work.

Letting W_i be the wage rate for the ith woman and using the notation of the text, woman i works and uses formal sources of child care if

$$X_i\beta + \varepsilon_i > 0 \text{ and } \ln W_i > \alpha_0 + \alpha_2 Y_i + \alpha_4 Z_i + u_i, \tag{A1}$$

where $X_i\beta + \varepsilon_i$ is the quality-adjusted price of child care. Woman i works and uses informal sources if

$$X_i\beta + \varepsilon_i < 0 \text{ and } \ln W_i > \alpha_0 + \alpha_1 X_i\beta + \alpha_2 Y_i + \alpha_4 Z_i + \alpha_1 \varepsilon_i + u_i. \tag{A2}$$

For nonworking women who would use formal sources,

$$X_i\beta + \varepsilon_i > 0 \text{ and } \ln W_i < \alpha_0 + \alpha_2 Y_i + \alpha_4 Z_i + u_i, \tag{A3}$$

while for nonworking women who would use informal sources

$$X_i\beta + \varepsilon_i < 0 \text{ and } \ln W_i < \alpha_0 + \alpha_1 X_i\beta + \alpha_2 Y_i + \alpha_4 Z_i + \alpha_1 \varepsilon_i + u_i \tag{A4}$$

To circumvent the practical difficulty that wage rates are missing for nonworking women, I postulate a wage function for the ith individual,

$$\ln W_i = m + nk_i + v_i, \tag{A5}$$

where k_i includes determinants of market wages such as education and labor-market experience, and where v_i is a disturbance term with zero mean and variance σ_v^2.

To derive the appropriate sample likelihood function, consider the joint density of the disturbances $h(u, v, \varepsilon)$ For a working woman who uses formal sources, the density function for u, v, and ε for the domain defined by (A1) is

$$\frac{h(u, v, \varepsilon)}{Pr(m + nk - \alpha_0 - \alpha_2 Y - \alpha_4 Z > u - v \wedge \varepsilon > -X\beta)}, \tag{A6}$$

where $Pr(\cdot)$ denotes the probability of event "$\cdot$". Assuming independence between ε and other disturbances, but allowing for dependence between u and v, we may write $h(u, v, \varepsilon) = g(u, v)s(\varepsilon)$, so that the conditional density of u, v, and ε, given that a woman works and uses formal sources, may be written as

$$\frac{g(u, v)s(\varepsilon)}{Pr(m + nk - \alpha_0 - \alpha_2 Y - \alpha_4 Z > u - v)Pr(\varepsilon > -X\beta)}. \tag{A7}$$

The density function for observed hours of work, wages, and the event "use of formal child-care services" is readily derived. To simplify the exposition, I present only the distribution for u, v, and the use of child care, which is

$$\frac{g(u, v)\int_{-X\beta}^{\infty} s(\varepsilon)\, d\varepsilon}{Pr(m + nk - \alpha_0 - \alpha_2 Y - \alpha_4 Z > u - v)Pr(\varepsilon > -X\beta)}. \tag{A8}$$

For working women who use informal sources, it is more convenient to work with the density function

$$b(u + \alpha_1 \varepsilon, v, \varepsilon), \tag{A9}$$

which can easily be derived from h. For expository ease, it is convenient to define $u + \alpha_1 \varepsilon = c$. Clearly, c is not independent of ε. This implies that the conditional distribution of c, v, and the dummy variable associated with the "use of informal sources" may be written as

$$\frac{\int_{-\infty}^{-X\beta} b(c, v, \varepsilon)\, d\varepsilon}{Pr(m + nk - \alpha_0 - \alpha_1 X\beta - \alpha_2 Y - \alpha_4 Z > c \Lambda \varepsilon < -X\beta)}. \tag{A10}$$

We may combine observations described by (A8) and (A10) to form the *conditional* likelihood function for working women. To incorporate sample information for nonworking women, we know that the appropriate probability statement for nonworking women who would not use formal sources is

$$Pr(m + nk - \alpha_0 - \alpha_1 X\beta - \alpha_2 Y - \alpha_4 Z < c \Lambda \varepsilon < -X\beta), \tag{A11}$$

while for nonworking women who would use formal sources the appropriate probability statement is

$$Pr(m + nk - \alpha_0 - \alpha_2 Y - \alpha_4 Z < u - v \Lambda \varepsilon > -X\beta). \tag{A12}$$

We may combine observations described by (A8) and (A10) with observations on nonworking women by weighting the conditional densities by the probability of the conditioning event.

To form the sample likelihood function in terms of observable variables, substitute for u and v from equations (5) and (A5), respectively. The Jacobian of transformation is $|\alpha_3 + \alpha_2 W(1 - e^{-\alpha_3 h})|$ It is straightforward to verify that since $h \geq 0$, the transformation is one-to-one since $\alpha_2 > 0$ and $\alpha_3 > 0$. Ordering the observations so that the first T_{WF} work and use formal sources, the next T_{WI} work and use informal sources, the subsequent T_{NWF} do not work and would use formal sources, while the last T_{NWI} do not work and would use informal sources, the sample likelihood may be written as

$$\begin{aligned}
= & \prod_{i=1}^{T_{WF}} |\alpha_3 + \alpha_2 W_i(1 - e^{-\alpha_3 h_i})| g(u_i, v_i) \int_{-X_i\beta}^{\infty} s(\varepsilon)\, d\varepsilon \\
& \prod_{i=T_{WF+1}}^{T_{WI}+T_{WF}} |\alpha_3 + \alpha_2 W_i(1 - e^{-\alpha_3 h_i})| \int_{-\infty}^{-X_i\beta} b(c_i, v_i, \varepsilon)\, d\varepsilon \\
& \prod_{i=T_{WF}+T_{WI}+1}^{T_{WF}+T_{WI}+T_{NWF}} Pr(m + nk_i - \alpha_0 - \alpha_2 Y_i - \alpha_4 Z_i < u - v) Pr(\varepsilon > -X_i\beta) \\
& \prod_{i=T_{WF}+T_{WI}+T_{NWF}+1}^{T_{WF}+T_{WI}+T_{NWF}+T_{NWI}} Pr(m + nk_i - \alpha_0 - \alpha_2 Y_i \\
& - \alpha_4 Z_i < c_i - v_i \Lambda \varepsilon < -X_i\beta)
\end{aligned}$$

where $u_i = \ln W_i - \alpha_0 - \alpha_3 h_i - \alpha_4 Z_i - \alpha_2[W_i h_i + Y_i - (W_i/\alpha_3)(1 - e^{-\alpha_3 h_i})]$, $c_i = u_i - \alpha_1 X_i \beta$, and $v_i = \ln W_i - m - nk_i$. In deriving the estimates, I assume that the disturbances are jointly normally distributed, with ε uncorrelated with u and v. The precise functional forms are well known and are not presented here. Note that each term in the third group of products involves the product of two cumulative normals and hence is easily computed. The last group of products involves a bivariate normal density. Rather than using inexact and computationally expensive numerical quadrature to evaluate these expressions, I use a Chebyshev-Hermite orthogonal polynomial expansion to 100 powers (Kendall 1941). This procedure allows for a computationally simple algorithm which can readily be applied to more general multivariate probability statements arising from the normal distribution.

To perform the calculations, two algorithms were used: the Powell conjugent gradient method and GRADX. The Powell method was used for initial iterations while GRADX was used to locate final optima.[22]

Appendix B

Tax Rates and Welfare Systems

In the text, I assume that the wage rate is independent of hours of work. My approach allows me to relax this assumption. Tax laws set rising rates of taxation on income intervals. For any wage rate, the income intervals can be converted into intervals for hours of work. Given an initial position for prework income, Y, the budget line is a broken line, illustrated in figure 6.

If the wage rate is known, we know the intervals t_1, t_2, and so on. Suppose a woman's equilibrium position for hours of work is C. Since the budget constraint is concave, the slope of the first segment of the constraint must exceed the slope of the indifference curve intersecting Y for a woman to work any hours. For a woman to work beyond Tt_1 hours, exactly the same type of inequality condition must be met. The "reservation wage" for work in excess of Tt_1 hours is partly determined by the value of hours worked and the known income YM which accrues if the woman works Tt_1 hours at the given wage rate net of taxes. Thus, a second inequality must be satisfied for a woman to work beyond Tt_1 hours. We acquire as many inequalities as there are breaks in the schedule. If a woman happens to be in equilibrium exactly at a switch point, the usual marginal conditions do not apply, but we can still acquire information about consumer preferences from such observations, since at the switch point, the marginal rate of substitution must lie between the slopes of the two segments of the budget constraint.

Since wage rates are known for working women, we can avoid bias due to censoring and incorporate tax rates into the estimation procedure by maximizing the appropriate conditional likelihood function.[23] To illustrate the procedure, consider a woman in equilibrium at position C in figure 6. For the woman to be in equilibrium on this segment of the budget constraint, the marginal rate of substitution at the right boundary is less than the slope of the segment, while the marginal rate of substitution at the left boundary is greater than the post tax marginal wage rate for further hours of work. This imposes a condition on the range of the disturbances of the indifference system. It is straightforward to derive the conditional distribution for the indifference system in this segment and to form the appropriate conditional likelihood function.

Precisely the same method may be applied to welfare systems which impose similar "distortions" on the budget line.

---

[22] For a discussion of these algorithms, see chap. 1 of the excellent book by Goldfeld and Quandt (1972).

[23] It is possible to have unknown wage rates, but the distributional problems become severe and they are not discussed here.

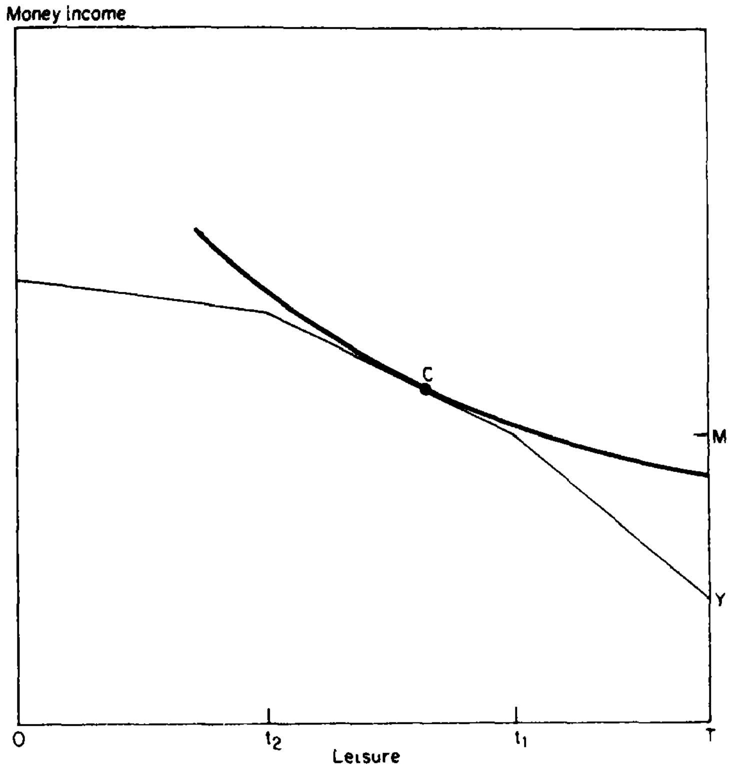


Fig 6

Comment

Sherwin Rosen

University of Rochester and Harvard University

Heckman has applied sophisticated theory and empirical methods to an important practical problem He uses a variant of Tobit analysis to combine qualitative data on labor-force status with quantitative data on hours of work. These methods not only take account of the usual differences in market wage opportunities confronting individuals, but also account for differences in tastes among them Few previous labor-supply studies have attempted to incorporate variations in tastes, and virtually none has treated the problem with other than ad hoc methods In addition, Heckman's statistical approach directly addresses a problem of interpretation raised by Lewis (1972): Cross-section regressions of labor-force participation rates on wage rates and other variables do not necessarily capture underlying substitution effects between leisure and goods. For example, suppose there are differences in preferences or home productivities among members of the population under study, over and above those represented by such exogenous variables as education and number of children Then there is a distribution of unobserved reservation wages Those whose reservation wage exceeds available market wages choose not to participate in the labor market. They specialize in home production and leisure instead. As wages increase, the fraction of the population satisfying this criterion decreases and labor-force participation rates rise Therefore, wage variation in the sample sweeps out a tail of the underlying distribution of reservation prices. Regression estimates may only give us back properties of this distribution and may be uninformative with respect to the more interesting aspects of the structure of preferences Heckman explicitly deals with this problem in his analysis. Furthermore both labor-force status and hours-of-work data are utilized efficiently i estimating labor-supply parameters. I know of no other work th utilizes the sample information in this way.

Heckman does not rest content at this point. He offers some conceptual modifications as well. Predictions of labor-supply responses to child-care programs require more knowledge than the uncompensated supply function, because these programs introduce a complicated wedge between average and marginal wage rates. However, no difficulties arise if the whole preference map or, what amounts to the same thing for prediction purposes, the utility compensated supply function for hours of work is known. The paper gives the impression that parameters of indifference curves are estimated directly. However, appearances are deceiving. Under the assumptions of the model, the plain fact is that the observations are generated by an uncompensated, garden-variety supply function. Conversely, the data identify an ordinary supply function, not an uncompensated one. The latter, as well as indifference-map parameters, always must be inferred from the former, using the theory of revealed preference. Since most of this theory is found in rather inaccessible discussions of integrability and related topics, it is worthwhile to spell out some details in a simple case.

A decision maker maximizes a quasi-concave utility function $u = U(x, h)$ subject to the constraint $x = wh + y$, where h is hours of work, x is all other goods, w is the wage rate (in terms of units of x), and y is nonlabor income (also in units of x).[1] Necessary conditions are $U_x = \lambda$ and $U_h = -w\lambda$, where λ is the marginal utility of income. These conditions and the budget constraint imply a set of uncompensated demand and supply functions,

$$x = f(w, y) \tag{1}$$

and

$$h = g(w, y). \tag{2}$$

It is equation (2) that generates the labor-supply observations. Equations (1) and (2) are not independent of each other. Substituting into the budget contraint and differentiating gives

$$wg_w - f_w = -g \tag{3}$$

and

$$-wg_y + f_y = 1. \tag{4}$$

Define an indirect utility function by substituting the demand and supply functions into $U(x, h)$: $u = U[f(w, y), g(w, y)] = V(w, y)$. Totally differentiate $V(w, y)$, $du = (U_x f_w + U_h g_w)dw + (U_x f_y + U_h g_y)dy$, and substitute the necessary conditions plus restrictions (3) and (4). Simplification yields the result

---

[1] Alternatively, $U(x, h)$ can be a synthetic utility function, after optimizing out the child-care quality decision conditional on h. That is the approach followed by Heckman. Then the quality price of child care is a proper argument of U, but is suppressed in this discussion.

$$du = \lambda[g(w, y)dw + dy]. \tag{5}$$

What combinations of wage rates and nonlabor income results in the same level of utility when consumption-labor choices are optimal?[2] Setting $du = 0$ in (5) and using the fact that $\lambda > 0$, these values must satisfy

$$dy/dw = -g(w, y). \tag{6}$$

Equation (6) defines the marginal rate of substitution between w and y It gives the slope of an "indifference curve" between w and y rather than between x and h Notice that the slope only depends on $h = g(w, y)$, which is known and estimated from the data.[3] Now treat equation (6) as a differential equation in y and integrate to obtain

$$y = F(w, c), \tag{7}$$

where c is a constant integration, dependent on the initial state (w_0, y_0). The conditions under which a solution exists are minimal and almost certain to be satisfied by any empirically tractable supply function In fact, (7) is the equation of the indirect preference map and c is a utility indicator. $F(w, c)$ is an inverse of the indirect utility function. Once the uncompensated supply function $h = g(w, y)$ has been estimated, $F(w\ c)$ can be obtained without prior knowledge of the utility function by integration, if not analytically, then with numerical methods on a computer Finally, substitute equation (7) into equation (2),

$$h = g[w, F(w, c)] = G(w, c) \tag{8}$$

Equation (8) is the utility-compensated supply function for hours of work and $G_w = g_w + g_y F_w = g_w - hg_y$ is the pure substitution effect Alternatively, solving for w in terms of h and c in equation (8) yields the marginal rate of substitution function, the form used by Heckman. Note that (8) or its inverse is inferred only from knowledge of the uncompensated supply function.

The discussion above shows how it is possible to learn something about the utility function purely from observing the uncompensated supply function. Heckman's procedure is more or less the opposite. He begins with a functional form restriction on the preference map (i.e., a restriction on equation [8]—his equation [4]) and integrates back to an uncompensated supply function (his equation [5], comparable with equation [2] above) It is the latter that is estimated. His restrictions are such that estimated parameters of the uncompensated supply function exactly identify parameters of the preference map without further manipulation Of course the theory does not specify precise functional forms, and some prior specifications are necessary for estimation The compensated and uncompensated supply functions stand in a one-to-one relationship, and a functional form restriction on either one implies a corresponding restriction on the other. Thus, the question boils down to where one should impose the restrictions in practice My own preference is to put the functional form restrictions on the uncompensated supply function and work back from that to the compensated function After all, the observations directly identify only the uncompensated supply function, and, in the nature of the revealed preference argument, the structure of preferences only can be inferred indirectly from observed behavior. There really is no way out of a specification search at the point of observations We simply have no way of knowing whether semilog linear preference maps are appropriate The proper specification must come from observed, uncompensated behavior.

---

[2] The following development is a variant of an argument due to Hotelling (1932)

[3] Differentiation of (6) reveals that $d^2y/dw^2 = -\partial h/\partial w + h(\partial h/\partial y)$. This is the negative of the pure substitution effect and the wage-nonlabor income indifference curves are concave to the origin.

Heckman's restriction makes the generating function of the observations (his equation [5]) nonlinear and difficult to estimate Computational difficulties are compounded by incorporating observations on labor-force nonparticipants. On the other hand, predictions of various program effects are more easily computed. If one starts from the other end, as suggested here, estimation is far less costly, and much greater experimental flexibility is achieved at the stage where that experimentation is most desirable. But this has its costs too, for preference-map inferences and program projections become more expensive. Still, if complex numerical integrations are necessary to obtain compensated response functions, they need be computed only once, after the functional form issue has been resolved Finally, Heckman's arguments in favor of his procedure are slightly more persuasive when income tax distortions are introduced into the analysis. However, it remains true that within the received framework the observations are generated by uncompensated structures Therefore, most of the above applies after average wage rates have been replaced by their appropriate after-tax marginal values.

Let me now turn to the empirical estimates If nonworking women are to provide information about the structure of preferences, it is necessary to know what market wage opportunities they have foregone Extrapolations are made from wages of working wives with the same productivity (education and experience) characteristics Yet there are theoretical reasons for supposing these extrapolations are biased, along the lines of the paper presented by Mincer and Polachek in this volume Women who have more-permanent labor-force attachments engage in productivity-enhancing activities to a much greater extent than those who have less permanent attachments.

To put it another way, "Are hours 'caused' by wages, or are wages 'caused' by hours?" Heckman's method assumed an answer of "no" to the second part of this question. But Mincer and Polachek's results suggest the opposite. Women who greatly value market work or who have relatively low home productivities have prior expectations leading them to acquire labor-market skills to a much different degree than women possessing the opposite characteristics. In practice, the result is that estimated market wage opportunities of those out of the labor force tend to be overstated. Estimates of substitution terms in labor-supply equations are biased upward, resulting in upward biased projections of labor-supply responses to child-care programs. Heckman has estimated the model separately for working wives, but the comparisons do not dispose of this point because the statistical treatment of child-care sources is quite different

My final comment concerns an anomaly in the data. The observations are very thin in the range $0 < h < 600$, and almost all women who engage in market activities work a significant amount of time during the year It is problematic how semilog linear reservation wage functions and normally distributed unobserved differences in preferences and market opportunities can account for this gap in the observations. If there is a significant mass of individuals at a boundary solution ($h = 0$) and the distributions are smooth, a nontrivial fraction exhibiting interior solutions close to the boundary is to be expected This raises some questions regarding sensitivity of the estimates to alternative distributional assumptions. Considering the estimation difficulties associated with normally distributed unobserved "error" components, pursuing other distributions would appear to be infertile ground for future research

But there is a much simpler explanation for the gap. There may be almost no jobs that offer short work schedules: Market wage opportunities in this range of hours may be nil. The point can be generalized. Most jobs outside the personal service sector offer rather inflexible work schedules (the major exception being volunteer work), with hours-of-work requirements varying from job to job. Insofar as choice of working hours is intermediated by a commensurate choice of job, there is no reason for a single hourly wage to clear all markets for jobs. That is, it may be factually incorrect to assume equality between average and marginal wage rates in the analysis. Wage-hours regressions are not readily interpreted as direct estimates of labor-supply functions in this case. They undoubtedly are related to worker preferences, but contain elements of employer demand and technological considerations as well. It is impossible to say at this point what difference such modifications will make to labor-supply estimates, since a slightly different conceptual framework is required. In any event, this issue seems well worth pursuing.

In conclusion, Heckman clearly has opened a lot of new territory in the economics of labor supply The originality of his effort is all the more remarkable when one considers the extent to which the literature shows these grounds have been worked over. Future studies in this area are bound to be affected for many years to come by Heckman's paper

Economics of Postwar Fertility in Japan: Differentials and Trends

---

Masanori Hashimoto

University of Washington and National Bureau of Economic Research

I. Introduction

Japan experienced a precipitous decline in fertility during the decade following the postwar baby boom of 1947–49, and then her fertility rates leveled off The rapid decline in fertility paralleled liberalization of abortion laws and an active campaign to disseminate contraceptive information Given these developments and the apparent subordination of women in Japanese culture, one might be inclined to ascribe the fertility decline to increased availability of abortions and other means of birth control and to doubt the explanatory relevance of economic theory which emphasizes the effects of rising wages and educational attainment of women on fertility. The conclusion that emerges from this study, however, is that the basic economic forces identified by the new economic theory of household decision making have been operating to produce a considerable part of the observed differentials and trends in Japanese fertility [1] Indeed, increased use of abortion and contraceptive devices appears to have been induced to some extent by economic forces [2]

I benefited greatly from comments and suggestions made by Gary S Becker, Margaret L Hashimoto, Masatoshi Kuratani, H Gregg Lewis, Gail Makinen, Jacob Mincer, Marc Nerlove, Douglas S Paauw, Gary Saxonhouse, T W. Schultz, T Paul Schultz, Miron Stano, Kozo Yamamura, and Dennis Zimmerman Haeng-Ja Song's research assistance is gratefully acknowledged Sachio Kohjima of the Bank of Tokyo obtained some data for me This paper was written while I was at Wayne State University I thank the Economics Department at Wayne State University, where this research was performed, for providing resources for my research Finally, I alone am responsible for the contents of this paper

[1] Silver (1966) found that births, and to a somewhat lesser extent marriages, responded positively to both ordinary and Kuznets cycles between 1878 and 1957, and that the magnitudes of the cyclical responses of births were similar to those in the United States and Britain

[2] It would be premature, however, to look for exact magnitudes of the parameters of the Japanese fertility function or for a complete explanation of fertility trends The statistical model employs only a few key variables and is still formulated in static terms The statistical analysis is performed mainly on cross-section data which are highly aggregated Further refinements of the model and the data base are both desirable and

II. Differentials and Trends in Japanese Fertility: Background

The phenomenal growth of the Japanese economy during the postwar years has been widely noted. Between 1950 and 1969, real gross national product grew at an average annually compounded rate of close to 10 percent, while the comparative rate for the United States was 4 percent. Real gross national product per capita grew at 9 percent, contrasted with 2 percent for the United States. The rapid industrialization of the Japanese economy is documented in the share of agricultural employment, which fell from about 49 percent in 1950 to 30 percent in 1960 and then declined to less than 18 percent in 1970. Employment opportunities for women grew also, the proportion of women in nonagricultural employment increased from 29 percent in 1950 to 35 percent in 1960 and to 37 percent in 1970.

Against this background of rapid economic development, Japanese fertility dropped sharply (fig. 1). The crude birth rate, which had been about 35 in the late 1940s, first dipped below 30 in 1950, 5 years later it fell below 20 and stayed there. Fertility decline in Japan began about 1920, but the rate of decline accelerated after 1950.[3] By the early 1960s the persistent decline appears to have ended, and then the leveling off began. A sharp drop in fertility in 1966 in response to the superstition of *Hinoeuma* vividly underscores that fertility is very much within the realm of deliberate control.[4] Birth rates just before and after 1966 were to some extent adjustments to the 1966 phenomenon.

Rural regions experienced a more rapid and a more persistent birth rate decline than industrial areas (see table 1). Only part of this difference is accounted for by shifts in age composition as women of childbearing age migrated from rural to urban regions.[5] Until 1962, the average number of children ever born by duration of marriage for women married before they were 30 years old continuously declined for all groups (see table 2). There was a slight increase between 1962 and 1967 among women married less than 10 years. But completed fertility measured approximately by the fertility of women married more than 10 years continued to decline. Table 3 provides more direct evidence on the decrease in desired completed fertility during the postwar years, 1950–67.

---

[3] Between the two periods, 1920–22 and 1940–42, the crude birth rate declined from 35.2 to 30.7 (7.1 percent per decade), and the total fertility rate declined from 5.21 in 1920 to 4.11 in 1940 (11.2 percent per decade). Between the 1949–51 and 1959–61 periods, the crude birth rate dropped from 28.8 to 17.2 (40.3 percent), and the total fertility rate from 3.72 to 1.99 (46.5 percent).

[4] *Hinoeuma* takes place every 60 years. The superstition is that a woman born in that year will "eat men alive," which is to say that she will be very aggressive and no man will want to marry her.

[5] According to Japan's Institute of Population Problems, holding constant the age composition of the population at 1930 values, the birth rate in urban areas (*shi*) dropped from 26.4 in 1950 to 17.8 in 1960 and then rose to 19.0 in 1965. The birth rate in rural areas (*gun*) continued to drop from 32.7 in 1950 to 19.0 in 1960 and to 17.6 in 1965. These results were reported in Aoki (1970, table 24).

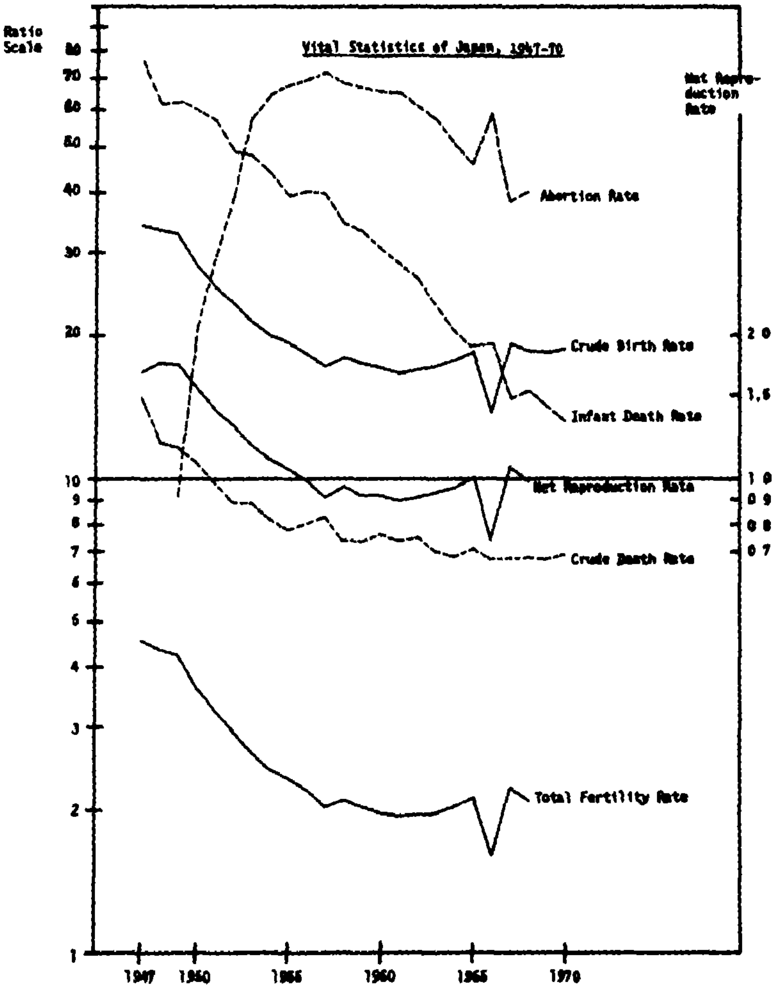


FIG 1.—Infant death rate is deaths at ages 0–1 per 1,000 live births Abortion rate is number of registered abortions per 100 live births Other vital statistics have their usual definitions. (Data on births, deaths, and infant deaths are from Japan Ministry of Health and Welfare [1970], abortion rate, total fertility rate, and net reproduction rate are from Mainichi Newspaper Company [1970, p 246]).

For families already having two children, the proportion desiring no additional children increased throughout the years, while the proportion desiring two and three or more additional children continued to decline Moreover, the percentage of couples with "no idea" how many additional children they wanted tended to decline, suggesting an increased

TABLE 1
Crude Birth Rates by Regions

	% Total Employment in Agriculture, 1960	Crude Birth Rates					% Change in Crude Birth Rates	
		1950	1955	1960	1965	1970	1950–60	1960–70
All Japan	30.0	28.1	19.4	17.2	18.6	18.8	−38.8	9.3
Industrial regions	11.7	25.6	16.8	16.8	20.1	20.7	−33.9	23.2
Rural regions	49.6	31.0	22.1	18.0	19.0	16.2	−42.6	−7.9

Sources.—Japan Ministry of Health and Welfare (1970, vol. 1) and Japan Prime Minister's Office (1970).

Note.—Industrial regions included are Tokyo, Kyoto, Osaka, Kanagawa, Fukuoka, and Aichi prefectures. Rural regions included are Aomori, Yamagata, Niigata, Ibaragi, and Kumamoto prefectures. Classification of prefectures as industrial or rural-farm is from Honda (1959). Crude birth rates for regions are simple averages of birth rates in each prefecture. Percent changes in crude birth rates for regions are simple averages of the percent changes in each prefecture.

TABLE 2
AVERAGE NUMBER OF CHILDREN EVER BORN
(WIFE MARRIED BEFORE AGE 30)

MARRIAGE DURATION (YEARS)	SURVEY YEARS			
	1940	1957	1962	1967
1–4	0.98	0.93	0.87	1.06
5–9	2.34	2.16	1.92	1.95
10–14	3.55	2.92	2.46	2.28
15–19	4.39	3.62	2.85	2.66
20+	5.17	4.72	3.90	3.43
All groups	3.49	2.86	2.37	2.28

SOURCES.—Aoki and Nakano (1967, table 3) and Japan Ministry of Health and Welfare (1968, table 24B).
NOTE.—Includes married women with husband present.

awareness of family planning in later years. My estimates of completed fertility (table 3) also show a pronounced decline over the years.[6]

The rapid decline in fertility in the postwar years coincided with the liberalization of policies on abortions and contraceptive use. Abortion was first legalized in September 1948, and in June 1949, grounds for abortion and sterilization were extended. After May 1952, abortion could be obtained with only the consent of man and wife. Alarmed at the increased number of abortions and possible health hazards from abortion, the government in 1952 initiated programs to disseminate birth control information through local agencies. Resort to abortion continued to increase, however, until 1957, when the number of abortions per 100 live births reached its peak of 71.6.

III. Economic Framework for Fertility Analysis

The following discussion will emphasize the effect of parental education on fertility. The testing of economic hypotheses about the effect of parental education on fertility also provides a test of the usefulness of economic theory in interpreting Japanese fertility behavior.

The basic proposition in the economic theory of fertility is that observed changes and differentials in fertility reflect to a large extent changes and differentials in the demand for fertility and that resource constraints on

---

[6] The true decline is likely to have been even more rapid than these estimates suggest because my calculations may underestimate completed fertility to a greater extent in earlier years. First, the true mean of the open-ended class would be larger than three, the figure used for my calculations, by a larger amount in earlier years if completed fertility declined over time. Second, the group with "no idea" as to number of additional children desired was omitted from the calculations. If people gave the "no idea" answer because of greater ignorance (lower level of education) than other respondents, then economic analysis predicts their desired completed fertility to be greater than that of others. This omission would have mattered less in recent years because of general declines in fertility and because this group made up a smaller fraction of respondents in more recent years.

TABLE 3

ADDITIONAL CHILDREN DESIRED AND ESTIMATED DESIRED COMPLETED FERTILITY, BY COUPLES WITH TWO CHILDREN

(WIFE'S AGE LESS THAN 50)

Number of Additional Children Desired	1950	1952	1955	1957	1959	1961	1963	1965	1967
0	29 8	39 3	42 7	56 4	57 7	64 2	71 7	70 5	71 1
1	32 8	35 2	32 3	30 0	25 5	19 3	22 3	23 5	20 0
2	19 2	14 7	15 8	6 9	7 2	4 0	3 2	2 4	3 1
3+	6 9	5 4	3 2	0 9	1 4	3 2	0 6	0 4	0 8
No idea	11 3	5 4	6 0	5 8	8 2	9 3	2 2	3 2	5 0
Estimated desired completed fertility	3 04	2 85	2 78	2 49	2 48	2 41	2 31	2 30	2 30

SOURCE —Aoki and Nakano (1967, p 44)

NOTE —Estimated desired *completed fertility* was calculated as 2 plus the average number of desired additional children Respondents with "no idea" were not included, and the 3+ group was assigned a value of 3 Figures are shown as percentages households exert appreciable effects on fertility demand Following Becker (1960), most economic models of fertility view children as primarily consumer durables. In this conception, a flow of services emanates from the stock of children, and these services enter the parents' utility function. Since parents must make outlays (e.g., food, clothing, housing, and, most important, their time) to obtain child services, they are not free goods, but command a price. To be quite general, therefore, demand for fertility is determined by the preferences of parents, the resources available to parents, and the cost of producing child services relative to the cost of other commodities (usually identified collectively as the standard of living) that also yield utility.[7]

Typically, the fertility rate has declined in countries experiencing sustained growth in real income Since growth in real income is usually brought about by growth in real wages, this phenomenon may be interpreted as reflecting the dominance of substitution over income effects An increase in real income would, other things being equal, increase the number of children per family unless the number of children was an "inferior good."[8] Assuming that child services are more time-intensive for mothers than the standard of living, an increase in the wife's wage rate would raise the relative price of child services, imparting a substitution effect away from them [9] Relying on income and wage rates alone, one may, therefore, interpret the decline in fertility accompanying growth in real income as reflecting the dominance of the substitution effect of a rise in the wife's wage rate over the combined income effects of increases in both spouses' wage rates. However, both income and wage rates are correlated with the educational levels of the parents and the level of contraceptive knowledge, relationships which complicate the empirical estimation of the magnitudes of the income and substitution effects

The effect of educational attainment of parents on fertility has been investigated more thoroughly than that of income or wage rates by Ben-Porath (1970*a*, 1970*b*, 1973) and Michael (1971) among others. Educational attainment of parents is typically negatively associated with their fertility. Proposed explanations of this relationship fall generally into three groups the tastes hypothesis, the cost-of-fertility-control hypothesis, and the cost-of-time hypothesis.

---

[7] Most models use the time-allocation framework (Becker 1965) and they are usually cast in a lifetime dimension, thereby avoiding the complexities of timing and spacing of births (see, for example, Ben-Porath [1973], DeTray [1972], Michael [1971], and Willis [1973]) Some problems with the use of static models are discussed by T. Paul Schultz (1973).

[8] Also relevant to the decline in fertility are changes in the quality of children as real income rises. Research on the demand for consumer durables suggests that the effect of income on the demand for quality is greater than on the demand for quantity. Even assuming numbers of children are not inferior, parents would tend to a greater extent to spend more per child than to have more children as real income rises.

[9] An increase in the husband's wage rate, however, is expected to produce mainly income effects because his time intensities are assumed to differ little between child services and the standard of living and because his wage income constitutes a large proportion of family income (see, for example, Ben-Porath [1973]).

According to one version of the tastes hypothesis, parents with higher education may be more "rational" than others in that they tend to regard the number of children as a decision variable, while less educated parents regard it as something beyond their control (Ben-Porath 1970*b*) Therefore, parents with higher levels of education would control fertility more and have fewer children than other parents Another version is that education affects the perception of child quality (Ben-Porath 1970*b*, Gardner 1972) Parents with more education may desire a greater expenditure per child to obtain a given level of child quality Other things being equal, they will have a smaller number of children. Economists have not succeeded in incorporating such taste factors in their theory in ways that generate testable hypotheses Such taste factors are still regarded as residuals

The second explanation hypothesizes that parents with more education have better access to information about birth control devices and are more efficient at using them. In other words, education lowers the cost of fertility control. More educated parents would engage in fertility control to a greater extent than other parents, other things, particularly the benefit of control, being the same. This hypothesis may partly explain the negative association between education and fertility [10]

The third explanation emphasizes the effect of education on fertility through its effects on the cost of time. In the time-allocation framework of the fertility model, education is assumed to raise the productivity of time spent at home in the production of commodities (such as child services and the standard of living) as well as that of time spent in the labor force. An increase in the productivity of time at work leads to an increased real wage rate, thereby raising the relative prices of time-intensive commodities. An increase in the productivity at home, say, at the same rate for all commodities, lowers the relative prices of time-intensive commodities. An increased productivity at home or market also results in increased real income, imparting income effects The full effect of education on the demand for children is the net of these forces To rely on the cost-of-time hypothesis to explain the negative correlation between education and fertility, one must argue that the substitution effect imparted by education outweighs the income effect. One plausible case is that education, especially of the wife, raises the productivity of time spent in the market more than that of time spent in producing child services or the standard of living. In this case, the relative price of child services rises, causing substitution effects away from them[11] The substitution effect of the husband's education is likely to be small, assuming that time intensity varies little between the two commodities for him The husband's education is expected, therefore, to exert mainly income effects (Ben-Porath 1973).

---

[10] More educated people may also use birth control more than others because they desire fewer children, i.e., the benefit of birth control is greater for them Both Becker (1960, p. 218) and Ben-Porath (1970*b*, p 6) propose a test of the hypothesis that education lowers the cost of fertility control. If improvement in birth control techniques comes as a once-and-for-all change, then fertility differentials among educational groups would widen at first but would tend to narrow over time as information on birth control measures became more diffused (see Section IV below for some evidence of a differential rate of diffusion among educational groups of birth control techniques in Japan after 1948–52)

The cost-of-time hypothesis has played a central role in the economic analysis of fertility It emphasizes the role of the relative price of child services, and its predictions are empirically distinguishable, at least for Japan, from those of the tastes and the cost-of-fertility-control hypotheses The latter two hypotheses predict that the education of either parent will be negatively associated with fertility, but they leave the relative importance of the effects of each parent's education an empirical question. In Japan, where the husband apparently dominates family decision-making,[12] one would expect his education to be at least as important as the wife's education, if not more so The cost-of-time hypothesis predicts that the negative association between education and fertility is more strongly manifested when the education variable refers to the wife's rather than the husband's education The husband's education would be positively correlated, if anything, with fertility when other things, in particular the level of contraceptive knowledge, are the same. In other words, husband's education serves as a proxy for family wealth, while wife's education represents the price of child services

These three hypotheses, however, are not necessarily mutually exclusive They help identify possible mechanisms through which parents' education affects fertility For example, the cost-of-time and the cost-of-fertility-control hypotheses would best be viewed not as alternatives but as complementary hypotheses The cost-of-time hypothesis emphasizes forces affecting the demand for the service of birth control, a demand which is derived from the desired reduction in births, given the level of fecundity. Underlying this demand is the demand for fertility. The cost-of-fertility-control hypothesis emphasizes the supply side, or the cost of achieving the desired reduction in births The cost-of-time hypothesis has been the most useful to economists for analyzing information on fertility behavior

---

[11] Even if the wife does not work in the market in her lifetime, her shadow price of time could be increased by education, making child services more expensive (see Gronau [1971])

[12] A survey of 422 couples in the Saitama prefecture in 1960 showed that 78 percent were using condoms, a device for the husband's use, and that 62 percent had the husband obtain the contraceptives. The same study showed that for 70 percent of white-collar workers, 52 percent of farmers, and 51 percent of laborers, the husband was responsible for obtaining contraceptive devices (Muramatsu 1962). Matsumoto, Koizumi, and Nohara (1972, p. 254) also report that husbands are primarily responsible for decisions concerning family limitation.

IV. Empirical Analysis

Study of Japan appears to offer an opportunity to obtain evidence concerning both the effects on fertility of decreases in the cost of preventing births and changes in fertility demand Thus, I want below first to examine evidence concerning the proposition that both decreased demand for fertility and increased access to abortions and contraceptive techniques contributed to the phenomenal postwar decline in Japanese fertility

Abortion, Contraceptive Use, and Fertility

The rapid decline in Japanese fertility in the postwar decade coincided, as I have stated, with abortion liberalization measures and the subsequent dissemination of information about other means of fertility control It is tempting to explain the rapid decline in fertility as simply the result of a widespread resort to abortion According to this view, many births prior to abortion legalization were unwanted the liberalization of birth-control policies simply eliminated this excess fertility In other words, if abortion had not been legalized, fertility would not have declined appreciably But since fertility in Japan had been declining persistently since around 1920, there is a presumption that it would have declined even in the absence of abortion reform

Since abortions and other fertility-control measures entail costs, both monetary and nonmonetary, the extent of use of these measures presumably was influenced by considerations of gains and costs Abortion legalization and birth control information programs certainly decreased the costs of fertility control, leading to increased use of control measures On the demand side, as discussed in Section III, the main determinant of the demand for fertility control is the demand for fertility A fuller interpretation of the developments in Japan would be that the price of child services increased and demand for fertility declined, leading to an increased demand for fertility control, and that the increased availability of birth control measures and abortions facilitated the attainment of lower fertility. Put differently, even without the liberalization of abortion and other control policies, resort to illegal abortion and the use of contraceptives would have grown because of the increase in the price of child services. In fact, the availability of illegal abortions had been growing before 1948, and the Eugenics Law of 1948 was motivated by hygienics, not by intent of population control (Muramatsu 1967) In short, legalization and liberalization of control measures were not so much autonomous events as responses by authorities to increased demand for fertility control and the health hazards of illegal practices.

How much did legalization of abortion contribute to the rapid decline in fertility? To answer this question accurately, one would have to estimate (1) the increase in the number of abortions caused by abortion legalization, and (2) the proportion of those abortions that would have terminated in births had there been no legal abortions. Since such estimates are beyond the scope of this paper,[13] I report the results of a crude calculation. Assuming that all pregnancies aborted legally would have terminated in live births if abortions had been illegal, I calculated, for each year from 1949 to 1961, the "maximum" general fertility rate as births plus abortions per 1,000 women of ages 10–49.[14] The excess of actual fertility decline over any decline in the "maximum" fertility rate can be viewed as a measure of fertility decline due to abortion.

Between the 1949–51 and 1959–61 periods, the actual general fertility rate dropped by 44 percent, from 93 to 52. According to my calculations, the "maximum" general fertility rate decreased by 22 percent, from 111 to 87. Thus, it would appear that abortions about doubled the rate of decline in the general fertility rate during the decade after 1950.[15] But resort to abortion by no means accounted for the entire fertility decline. Indeed, the rate of decline in the estimated "maximum" rate is still substantially greater than the rate of decline in the actual fertility rate during the prewar years.[16] This evidence of a substantial fertility decline—indeed, a possible acceleration of the prewar decline—after adjusting for the effect of abortions suggests that the actual fertility decline partly represents decreases in the demand for fertility.

---

[13] In principle, effects of legal abortions on fertility in any year could be assessed from a simultaneous-equations model in which key endogenous variables would be the fertility rate and prices of abortions and of contraceptive measures. Once the effect of legalization on the prices of abortions and contraceptive measures was estimated, the ultimate effect on fertility could be obtained from the demand-for-fertility function.

[14] These are maximum estimates for each year of the fertility rate in the absence of legal abortions because, if abortions had been illegal, some pregnancies would have been aborted illegally or prevented by other means. In other words, the abortions in each year were not solely due to decreases in costs but also to increases in the demand for fertility control. On the other hand, these estimates may understate the true maximum values because of underreporting of abortions. Muramatsu (1960) estimated that in 1955 at most 77 percent of induced abortions performed were reported. He also estimated that in 1955 the number of live births in the absence of abortion would have been twice (or more) the number actually registered. I calculated the ratio of the maximum to the actual general fertility rate to be 1.68 in 1955.

[15] Because of the underreporting of abortions, my calculations may overstate the effect of abortions on fertility decline. Since the underreporting was probably greater in earlier years, the actual drop in the maximum fertility rates would have been greater, and the effect of abortions on fertility decline smaller than I calculate.

[16] Between 1920 and 1940, the rate of decline per decade in the general fertility rate was about 13 percent.

TABLE 4
RATE AND COMPOSITION OF PREGNANCIES BY REGIONS, 1955, 1960, AND 1965

				RELATIVE CHANGE		AVERAGE 5-YEAR-PERIOD CHANGE (%)
	1955	1960	1965	1955–60 (%)	1960–65 (%)	
Rate of pregnancy						
All Japan	223 32	187 01	166 01	−16 26	−11 23	−13 74
Industrial regions	206 25	191 23	183 00	−7 28	−4 30	−5 80
Rural-farm regions	251 14	191 30	155 59	−23 83	−18 67	−21 29
Share of births (in %)						
All Japan	57 96	58 14	66 04	0 31	13 59	6 74
Industrial regions	54 59	56 79	65 91	4 03	16 06	9 88
Rural-farm regions	59 58	59 85	65 28	0 45	9 07	4 67
Share of abortions (in %)						
All Japan	39 17	38 49	30 53	−1 74	20 68	−11 71
Industrial regions	42 24	39 40	30 42	−6 72	−22 79	−15 14
Rural-farm regions	37 79	37 01	32 22	−2 06	−12 94	−7 66
Percentage of females (Ages 15+) of childbearing age (20–34) (in %)						
All Japan	36 88	35 73	34 21	−3 12	−4 25	−3 69
Industrial regions	39 28	38 95	39 29	−0 84	0 87	0 01
Rural-farm regions	35 66	34 06	30 17	−4 49	−11 42	−8 02

SOURCES —Pregnancies, births, and abortions from Shinozaki (1971), female population composition from Japan Prime Minister's Office (1955, 1960*b*, 1965)

NOTE —Rate of Pregnancy = 1000 × (births + fetal deaths + abortions)/married women of childbearing age with husband present Share of births = (births/pregnancies) × 100 Share of abortions = (abortions/pregnancies) × 100 Regions are the same as in table 1

Dissemination of birth-control measures other than abortions undoubtedly played some role in the decline in actual fertility rates, but in the absence of estimates of the number of pregnancies prevented by the use of modern contraceptive measures, one cannot directly assess its impact on fertility. Since the efforts to disseminate birth-control information started only in 1952, the major effects of birth control on fertility were probably felt during the late 1950s Assuming that most couples with accidental pregnancies turned to abortion as a last resort to control fertility, the share of abortions compared with the share of actual births in total pregnancies may provide information on the trend of fertility demand and the use of contraceptives Since abortions had already become widespread by the middle of the 1950s, subsequent changes in the share of abortions would reflect, to a large extent, an increased use of contraceptives Table 4 summarizes trends and differentials in the rate of pregnancy and in the shares of abortions and births in pregnancies after 1955. There was a general increase of the share of births and a decrease of the share of abortions in total pregnancies For Japan as a whole, the pregnancy rate among fertile women declined at an average 5-year rate of 14 percent from 1955 to 1965 The rate of decline was greater in the rural regions (21 percent) than in the industrial regions (6 percent) Indeed, by 1960 the pregnancy rate in rural regions had dropped below that in industrial regions. During the 10-year period, the rate of increase in the share of births and the rate of decrease in the share of abortions were both greater in industrial than in rural regions, suggesting that "unwanted" pregnancies declined more rapidly in industrial areas

What can one infer from these regional differences? Suppose that all these differences are attributable to the differential spread of contraceptive methods (i e , differential reduction in costs of control) Then the implicit assumption would be either that there was no change in demand for fertility in industrial and rural regions, or that the decline in demand for fertility did not vary systematically between the two types of regions The slower rate of increase in the share of births and the slower rate of decrease in the share of abortions in rural regions would then presumably be attributable mostly to a slower spread of contraceptive measures in these regions The more rapid decline of the pregnancy rate in rural regions, however, is the opposite of what one would expect if contraceptive measure did spread more slowly in these regions. Therefore, a more plausible interpretation of these differences would rely on a combination of differential change in fertility demand and differential spread of contraceptive knowledge One interpretation would be that fertility demand declined more rapidly in rural than in industrial regions, resulting in a more rapid decline of the pregnancy rate, but that the spread of contraceptive knowledge was not rapid enough in rural regions to diminish rapidly the use of abortions as a last resort to control fertility

Although the changing age composition of females precludes a firm judgment about the validity of the above interpretation,[17] this interpretation is supported by a comparison of the 1955–60 and 1960–65 periods The pregnancy rate dropped more rapidly during 1955–60 than it did during 1960–65 in both regions If this was attributable entirely to the spread of contraceptive measures, one would also expect to observe more rapid changes in both the share of births and the share of abortions in the earlier period Indeed, given the early 1950s' reforms regarding the use of contraceptives, this expectation would appear to be quite plausible, but the course of events was just the opposite as both shares changed more rapidly in the later period (table 4). Again, a more plausible interpretation would be that demand for fertility continued to decline in both periods; that the spread of contraceptives increasingly facilitated effective

---

[17] The industrial prefectures experienced a slight increase and the rural prefectures a rather large decrease in the proportion of women of childbearing age (table 4) These changes in age composition undoubtedly explain in part the more rapid decline in the pregnancy rate in the rural than in the industrial prefectures.

TABLE 5
DIFFUSION OF CONTRACEPTIVE USE BY EDUCATION LEVEL
(PERCENTAGE OF COUPLES USING CONTRACEPTIVE DEVICES)

	1950	1952	1955	1957	1959
All	19 5	26 3	33 6	39 2	42 5
Husband's education					
<9	14 2	18 2	28 2	33 4	37 6
10–12	25 4	37 0	37 7	46 5	43 9
13+	37 3	47 0	48 8	52 5	54 0
Range	23 1	28 8	20 6	19 1	16 4
(MAD/mean)	30 31	31 03	18 39	16 25	13 06
Wife's education					
<9	13 0	20 1	28 2	33 3	35 0
10–12	32 4	38 7	46 1	48 1	51 6
13+	36 0	59 1	47 8	53 2	51 9
Range	23 0	29 0	19 6	19 9	16 9
(MAD/mean)	34 75	33 59	20 47	17 16	16 09

SOURCE —Japan Ministry of Health and Welfare (1959)

NOTE —Range is the difference between the largest and the smallest values The MAD is the mean absolute deviation In all years except 1959 the data refer to couples with wife under age 49 For 1959, wife's age is less than 50 The sample size varies from 3,500 couples in 1952 to 3,835 couples in 1959

fertility control, thereby reducing unwanted pregnancies, and that the role of abortion continued to diminish in importance as a control measure

The above evidence documents an increased use of contraceptive measures from 1955 to 1965 The relationship between parents' education and the use of contraceptives was discussed in Section III in the form of the cost-of-fertility-control hypothesis If there was a period in which the use of birth control techniques was introduced or improved once and for all, then this hypothesis predicts that differences among education groups in the extent of use of birth control would widen at first, then narrow gradually In Japan, a marked increase in the availability of abortions and birth control information occurred after 1948–52, so this period would appear to provide a reasonably proper setting for testing this hypothesis

Table 5 presents the percentage of families within education class (husband's or wife's) that reported using some birth-control technique The data are taken from surveys conducted in selected years to determine the attitudes of Japanese families of various socioeconomic classes toward contraceptive usage In most education classes, the proportion of families using birth-control techniques increased over the years, presumably because costs of control declined and benefits increased The differences across education groups in the extent of contraceptive usage as measured by either the range or the mean absolute deviation (relative to the mean) varied most in the early 1950s and persistently declined after 1952 One detects a slight increase in these measures of dispersion in 1952 with the

TABLE 6
COMPLETED FERTILITY BY EDUCATION LEVEL
(AVERAGE NUMBER OF CHILDREN EVER BORN PER COUPLE)

	1940	1952	1962	1967
All	5.10	4.47	3.91	3.36
Husband's education				
Low	5.19	4.62	4.05	3.45
Middle	4.81	3.62	3.60	3.27
High	4.17	3.47	3.21	2.92
Wife's education				
Low	5.19	4.57	4.04	3.48
Middle	4.39	3.58	3.47	3.14
High	4.74	3.13	3.09	2.69

SOURCES.—Aoki and Nakano (1967, table 11) and Japan Ministry of Health and Welfare (1968, various tables).
NOTE.—For 1940, the figure is for couples married more than 21 years, for 1952, couples with wife age 45 or older, and for 1962 and 1967, couples married more than 20 years. Schooling definitions are low, less than 10 years, middle, 10–12 years, and high, 13 years or more.

start of the campaign to disseminate birth-control information. These findings, though based on sparse information, agree with the predictions of the cost-of-fertility-control hypothesis.

Education and Fertility

The preceding discussion suggests that a decline in fertility demand and increased access to abortions and birth-control information played a joint role in producing the observed decrease in fertility in Japan. I now explore the relationship between fertility and parental education in the light of the hypotheses discussed in Section III.

According to the tastes and cost-of-fertility-control hypotheses, education of either parent is expected to be negatively associated with fertility. For Japan, education of the husband, because of his dominant family role, is likely to be a more significant explanatory variable than is the wife's education. The cost-of-time hypothesis predicts that the husband's education will be weakly, but positively, associated with fertility and that the wife's education will have a strong negative association with fertility.

Table 6 presents a measure of completed fertility by education groups for 1940, 1952, 1962, and 1967. The data are from Institute of Population Problems surveys. Education of either parent generally is negatively associated with fertility, in agreement with findings for other countries. These data also document the persistently declining trend in completed fertility at all levels of parental education, underscoring that education is by no means the only factor contributing to the decline in completed fertility over time.

TABLE 7
TRENDS IN COMPLETED FERTILITY STANDARDIZED BY EDUCATION

	SURVEY YEARS				MEAN ABSOLUTE DEVIATION
	1940	1952	1962	1967	
Actual .	5 10	4 47	3 91	3 36	0 85
Standardized.					
Husband's education	4 99	4 21	3 84	3 36	0 75
Wife's education	4 93	4 22	3 84	3 36	0 72

SOURCE —Calculated from table 6
NOTE —Standardized rates were calculated by using the 1967 distribution of couples by education of husband and wife The mean absolute deviation was calculated as the average of absolute deviations of annual values from the 1967 value

The negative association of fertility and the education of either parent may at first appear to agree with the prediction of the tastes and the cost-of-fertility-control hypotheses Unfortunately, these data are not cross-classified by education of parents, so that one cannot discern the "pure" relationship between the fertility rate and the wife's and husband's education separately Since the levels of husband's and wife's education are (positively) correlated, table 6 reflects the joint influence of both levels of education. Since the correlation is not perfect, one may still attempt (see table 7) to determine the relative importance of the education of either parent in "explaining" observed differences in fertility

Table 7 presents the observed completed fertility in the 4 survey years and estimates of completed fertility obtained by using the 1967 distribution of educational attainment of husband and wife That the general increase in the level of school attainment played a role in the decline in observed fertility is demonstrated by standardized rates which are consistently lower than actual rates in every year The difference between standardized fertility rates using husband's and wife's education appears to narrow in later years One reason for this tendency may be the increased correlation of husband's and wife's educational attainment accompanying a general increase in the dispersion and level of women's educational attainment According to the mean absolute deviation of yearly values from the 1967 value (shown in table 7), the year-by-year differences in the fertility rate are decreased when yearly rates are standardized by education of either parent, but the reduction is somewhat greater when the distribution of the wife's rather than of the husband's education is used Standardizing by the husband's education reduces the mean absolute deviation by 12.8 percent, standardizing by the wife's education decreases it by 15.3 percent. Thus, the trend in the distribution of the wife's education appears to have been more important than that of the husband's in the fertility trend.

Since the data for table 7 were not cross-classified by parents' education, one could not directly observe the "pure" effect of the education of either parent To determine the relative magnitudes of the pure effects of parents' education on fertility, multiple regressions of the following form were estimated using cross-section data for 46 prefectures: $C = a + b_1(Ed_{HN}) + b_2(Ed_{HH}) + c_1(Ed_{WN}) + c_2(Ed_{WH}) + f(\text{other variables}) + u$, where C = fertility rate; Ed_{HN} = proportion of males who never attended school; Ed_{HH} = proportion of males who completed approximately 10 or more years of schooling, Ed_{WN} = the same as Ed_{HN}, for females, Ed_{WH} = the same as Ed_{HH}, for females; other variables = other standardizing variables discussed in the text, and u = disturbance term Since time-series data on education and other variables are lacking for Japan, as for many other countries, the regression analysis was restricted to cross-section data. Regressions reported in table 8 use the average number of children ever born for selected age groups of women (completed fertility) as the dependent variable, in regressions in table 9 the live birth rate is the dependent variable The Japanese census presents the distribution of educational attainment in four categories—none, elementary, middle, and high Instead of estimating a single measure of educational attainment, such as the mean or median, I used two categories of the education variable for each parent One advantage in using these measures is that the coefficients, that is, b_1 and b_2 for the husband and c_1 and c_2 for the wife, may reveal differential effects on fertility of a shift in the education distribution between the lower and the upper segments of the distribution The expected signs of the regression coefficients according to the hypotheses discussed earlier are shown below

Tastes and Cost of Fertility Control*		Cost of Time†	
$b_1 > 0$	$b_2 < 0$	$b_1 \leq 0$	$b_2 \geq 0$
$c_1 > 0$	$c_2 < 0$	$c_1 > 0$	$c_2 < 0$

* A weaker version of the above $-b_1 + b_2 < 0$ and $-c_1 + c_2 < 0$

† A weaker version of the above $-b_1 + b_2 \geq 0$ and $-c_1 + c_2 < 0$

Regressions of Completed Fertility

For the regressions of completed fertility reported in table 8, each observation was weighted by the square root of the number of age-specific ever-married women in each prefecture In addition to the education variables, standardizing variables included in these regressions are the percentage of persons, age 15 and older, employed in agriculture, forestry, and fisheries, and the percentage of the population living in urban areas These variables refer to 1960 and therefore do not reflect accurately the background in which the couples made fertility decisions For these variables to have substantive meaning, one must assume that the majority of couples have lived in the prefecture they resided in in 1960 for most of their married lives and that the rankings of prefectures by these variables have not changed significantly over time Unfortunately,

TABLE 8
REGRESSIONS OF COMPLETED FERTILITY, 46 PREFECTURES IN JAPAN 1960 CENSUS (WEIGHTED REGRESSIONS)

REGRESSION NUMBER	AGE COHORT OF WIFE	% EMPLOYED		EDUCATION (MALE)		EDUCATION (FEMALE)		$\bar{R}^2$/S E
		Agriculture	Urban	None	High	None	High	
1	(40–49) (n = 46)	0.006	−0.010	−0.179	−0.011	0.191	0.019	.938
		(0.71)	(−1.30)	(−1.07)	(−0.52)	(1.87)	(0.21)	.297
		F(2, 39) = 5.64*		F(2, 39) = 0.64		F(2, 39) = 2.05		
2	(50–59) (n = 46)	0.009	−0.009	−0.378*	0.006	0.213*	−0.099	.945
		(1.07)	(−1.16)	(−2.95)	(0.21)	(4.60)	(−0.84)	.325
		F(2, 39) = 7.96*		F(2, 39) = 5.03*		F(2, 39) = 11.38*		
3	Pooled (n = 92)	0.006	−0.011	−0.307*	−0.001	0.209*	−0.043	.941
		(1.16)	(−2.01)	(−3.17)	(−0.06)	(5.25)	(−0.62)	.608
		F(2, 84) = 14.18*		F(2, 84) = 5.18*		F(2, 84) = 14.19*		
4	Pooled	0.004	−0.012	−0.311*	−0.004	0.201*	−0.038	.942
		(0.51)	(−1.62)	(−3.15)	(−0.22)	(4.85)	(−0.55)	.612
		0.009	−0.010					
		(1.16)	(−1.17)					
		F(4, 82) = 7.28*		F(2, 82) = 5.03*		F(2, 82) = 12.12*		

SOURCE.—Japan Prime Minister's Office (1960*b* various volumes.)

NOTE.—Student's *t* values are in parentheses. Intercepts are not reported. Regression 3 allows only the intercept to vary between the two cohorts, while 4 allows the regression coefficients of "agriculture" and "urban" as well as the intercept to vary between the two cohorts. The dependent variable (average number of children ever born) and the education variables are age-specific.

* Significant at 1%.

TABLE 9

Regressions of Live Births per 1,000 Married Women, Ages 15+ (46 Prefectures in Japan 1960 Census)

Regression Number	Age Composition	% Employed		Education (Male)		Education (Female)		Log of Monthly Earnings		Labor-Force Participation Rate		$\bar{R}^2$/S.E.
		Agriculture	Urban	None	High	None	High	Male	Female	Male	Female	
1	2.765** (8.34)	0.189 (1.35)	−0.184 (−1.29)	0.647 (0.17)	1.729 (1.27)	3.279* (2.34)	−6.322 (−1.67)					0.739 5.411
		F(2, 38) = 3.29*		F(2, 38) = 0.83		F(2, 38) = 4.30*						
2	2.339** (6.78)	0.148 (1.04)	−0.039 (−0.27)	−5.287 (−1.31)	−0.565 (−0.38)	5.008** (3.26)	−0.659 (−0.16)	13.422 (1.28)	−9.624** (−2.74)			0.790 4.981
		F(2, 36) = 0.82		F(2, 36) = 0.84		F(2, 36) = 6.05**						
3	2.618** (6.12)	0.262 (1.35)	−0.138 (−0.88)	−0.855 (−0.20)	1.518 (1.03)	3.743* (2.35)	−6.111 (−1.58)			−0.063 (−0.07)	−0.189 (−1.90)	0.744 5.502
		F(2, 36) = 2.33		F(2, 36) = 0.57		F(2, 36) = 4.51*						
4	2.325** (5.74)	0.158 (0.87)	0.021 (0.14)	−7.208 (−1.59)	−1.131 (−0.69)	5.430** (3.33)	−0.263 (−0.06)	13.799 (1.23)	−10.045** (−2.87)	−0.635 (−0.72)	−0.062 (−0.54)	0.797 5.043
		F(2, 34) = 0.40		F(2, 34) = 1.30		F(2, 34) = 6.20**						

Sources.—Earnings Japan Prime Minister's Office (1960a) Other variables Japan Prime Minister's Office (1960b, various volumes)

Note.—Student's t-values are in parentheses. Age-composition variable is birth-rate-constant age distribution of women calculated by

$$\sum_{i=1}^{9} B_i W_{ij},$$

where B_i is the birth rate in the ith age class for all women in Japan, W_{ij} is the proportion of women in the ith age class in the jth prefecture (i = 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55+) Monthly earnings are average monthly contract wages and salaries for household head (male) and average monthly contract wages and salaries for wife and other household members (female) in 1959, each is deflated by a regional price index. Labor-force participation rates are percentages of populations ages 15+ in the labor force in 1960

* $p < .05$

** $p < .01$

neither income nor earnings is held constant in these regressions for lack of data pertaining to these age groups

Regressions 1 and 2 were run for married women, ages 40–49 and 50–59, respectively No coefficient in regression 1 is significant The ambiguous result obtained for this cohort may be partly because the majority of these women were married during World War II times Uncertainty about the future and participation by males in war activities may have distorted family-size decision making The sizes of the regression coefficients for the education variables satisfy the weaker prediction of the cost-of-time hypothesis, however For regression 2, the results are less ambiguous Judging from the t-values, both Ed_{HN} and Ed_{WN} are significant at the 1 percent level, and the F-values for the education variables indicate that both the husband's and the wife's education are statistically significant at better than 1 percent Both husband's and wife's education obtain regression coefficients that are largely consistent with the cost-of-time hypothesis

To determine if the fertility relationships for the two groups of married women were different, I tested three null hypotheses using Chow's (1960) test for equality of two regressions The three null hypotheses were that (1) the two relationships are completely the same, (2) the two relationships are the same except for the intercept, and (3) the two relationships are the same only with respect to the education variables The first null hypothesis was rejected ($F_{7,78}$ = 10 93), but the latter two were accepted Based on these findings, I pooled the observations for the two groups and obtained regressions 3 and 4 Regression 3 assumes that only the intercept differs, while regression 4 assumes that everything except the education coefficients differs between the two groups The results for pooled regressions 3 and 4 are quite similar Both husband's and wife's education are statistically significant at better than 1 percent, with regression coefficients as expected from the cost-of-time hypothesis

The results of these regressions clearly support the economic framework based on the cost-of-time hypothesis According to the coefficients of the pooled regressions, male education variables higher by one standard deviation unit, holding female education constant, are associated with higher fertility by about 0 13 children Holding male education constant, female education variables higher by one standard deviation are associated with about 0 30 fewer children However, the relationship between education and fertility is apparently nonlinear The distribution of either parent's educational attainment has a greater effect on average completed fertility when it shifts from no education to some education than when it shifts from some to high education [18] While influences of education on fertility through tastes and the cost of fertility control are undoubtedly present, the dominant influences appear to be through income and prices

---

[18] Ben-Porath (1973) reported a similar finding for Israel In particular, the sharpest decline in fertility was observed between women with no education and women with some education

These regressions refer to families which made fertility decisions before and during World War II. One would expect the decision-making process to have been more male-dominated in these than in postwar years Yet female education is one of the most important variables in these regressions, a finding which strongly supports the relevance of the economic analysis of fertility For this paper, regressions of completed fertility could not be run for families making fertility decisions during the postwar years, since the 1970 census was not available for all prefectures As an alternative, I now turn to regressions of live birth rates in 1960

Regressions of Live-Birth Rates

Multiple regressions of the number of live births in 1960 per thousand married women, age 15 or older, are reported in table 9 In addition to the percentage employed in agriculture and the percentage living in urban areas, these regressions hold constant the age composition of women and, in some cases, wage rates and labor-force participation rates The age-composition variable is a weighted average of age-specific birth rates for Japan, the weights being the percentage of women in each age class in each prefecture The percentage distributions of all women rather than of only married women were used because the latter may reflect endogenous decisions about the timing and duration of marriage Both the female wage rate and the female labor-force participation rate are expected to affect fertility negatively The male wage rate and the male labor-force participation rate would capture income effects and are expected, therefore, to obtain positive regression coefficients.

The results for the birth-rate regressions are largely consistent with the predictions of the cost-of-time hypothesis An increase in male education, holding constant female education, increases, and an increase in female education, holding constant male education, decreases the birth rate Female education is always more significant than male education Female education tends to remain statistically significant with negative effects on fertility even after the female wage rate and labor-force participation rate, either separately or together, are introduced in the regressions As expected, the female wage rate obtains negative and significant regression coefficients The male wage rate obtains coefficients that are insignificant, though with expected signs Neither the agricultural employment nor the urbanization variable has significant coefficients, perhaps because urban-rural differences in birth rates had largely disappeared by 1960 (see table 1).

Adding wage-rate variables to the regressions greatly reduces the effect of female education on the birth rate Without the wage-rate variables in regressions 1 and 3, a uniform percentage shift upward in the female education distribution reduces the birth rate by about 10, for example, $-3\ 279 - 6\ 322 = -9\ 6$ for regression 1, but the magnitude drops to about five when the wage-rate variables are introduced in regressions 2 and 4. This finding is not surprising since the cost-of-time effects are likely to be captured jointly by female education and wage rates The female wage rate also reduces the magnitude and significance of the female labor-force participation rate, as expected if decisions about fertility and labor supply are made simultaneously on the basis of the wage rate [19] The nonlinearity usually observed in the effects of female education is observed only when wage rates are held constant Otherwise, the results are ambiguous the deterrent effect of female education appears greater when education increases from some to high than from none to some schooling, though the regression coefficients for "high" are never statistically significant

These regressions suggest that area differences in the birth rate reflect differences in the female variables more than in the male variables For example, according to regression 2, increasing all the male variables by one standard deviation unit, holding constant other variables, increases births per thousand married women by about three, but similar increases in female variables decrease the birth rate by 10 Calculations for regression 4 show the effect of male variables to be $+1\ 2$ and the effect of female variables to be $-10\ 9$

V. Summary and Agenda

Increased access to abortions in Japan undoubtedly contributed to the post–World War II acceleration of that country's decline in fertility But I conclude that fertility would still have declined at a more rapid rate than in the prewar period, which had been characterized by a mild and rather monotonic decline, if abortion had not been legalized Therefore, dissemination of contraceptive information and/or an accelerated decline in the demand for fertility must also have contributed to the trend.

---

[19] A rise in the female wage rate, other things being equal, increases the attractiveness of labor-force participation relative to home production, including child rearing Therefore, increased female wage rates not only directly reduce fertility demand but also increase labor-force participation, which in turn indirectly reduces fertility still more Moreover, the direction of causation may also run from higher birth rates in an area to lower female labor-force participation there To allow for such simultaneity, regressions 3 and 4 were also estimated by two-stage least squares, treating female labor-force participation as endogenous However, these estimates did not substantially differ from the ordinary least-squares estimates The labor-force participation regressions seem to have been plagued by multicollinearity Although the regressions were statistically significant, judging by the F-values, with coefficients of determination above 0 70, they did not successfully disentangle the effects of the exogenous variables

Cross-section analyses of 1960 data reveal that the economic theory of fertility, with its emphasis on the effects of female education and earning power on fertility demand, provides a useful framework for analyzing fertility differentials in Japan. In particular, regressions of completed fertility and live birth rates in 1960 show that female education and earnings affect fertility negatively and are the most significant variables in the regressions Male education and earnings tend to have positive but less significant effects On the basis of only sociological, demographic, or taste considerations, one would not necessarily expect to find female education more important than male education in explaining fertility differentials. Indeed, in Japan, because of the male's generally dominant role in household decisions, one would have expected to find male education to be at least as relevant as female education, if not more so, in explaining fertility differences

How useful are these results in interpreting the fertility trends in Japan? A tentative answer can be obtained from the regressions in table 9 Assuming that the parameter estimates remain stable over time and using the observed magnitudes of change between 1950 and 1960 for all the explanatory variables except for the age-composition variable, my results predict the decline of births per thousand married women to be only 4 percent at the mean according to regression 2, and 3 percent according to regression 4 The actual decline between 1950 and 1960 was about 42 percent, of which one would hope to explain about half if legalization of abortion explains another half For the period 1960–70, both regressions predict a decline, while in fact births per thousand married women increased somewhat On the basis of these results, one may be inclined to doubt the usefulness of cross-section regressions in interpreting the trend [20]

However, since the regression models and the economic hypotheses underlying them pertain to long-run adjustments, one would expect from these regressions better predictions of longer-term changes in fertility It is possible that the sharp decline of fertility during the 1950s and the subsequent leveling off in the 1960s represent swings around a longer-term trend that has persisted since 1920 [21] If so, a more appropriate comparison of actual and predicted declines might be made between 1950 and 1970 The predicted decline for this period was 19 percent according to regression 2 and 15 percent according to regression 4. The actual decline was about 30 percent. These regressions appear, therefore, to be more helpful in providing an interpretation of the longer-term fertility

---

[20] T Paul Schultz (1973) discusses problems in the interpretation of time series from cross-section estimates of parameters Here I ignore the kinds of problems he raised.

[21] A similar observation is made by Okazaki (1970, p 26)

TABLE 10
A COMPARISON OF TRENDS IN FERTILITY AND ECONOMIC VARIABLES 1950–69 (JAPAN AND UNITED STATES)

	1950	1960	1969	RATES OF CHANGE 1950–60 (%)	 1960–69 (%)	 1950–69 (%)
Crude birth rate						
Japan	28 1	17 2	17 9*	−38 8	4 1	−36 3
U S (white)	23 5†	23 2‡	16 6	−1 3	−28 5	−29 4
Total fertility rate						
Japan	3 63	1 99	2 10§	−45 2	1 05	−44 6
U S (white)	3 05†	3 56‡	2 37‖	16 7	−33 4	−22 3
Real GNP per capita (1950 = 100)						
Japan	100	247 4	529 3	147 4	114 0	429 3
U S	100	115 2	152 7	15 2	32 5	52 7
Real wage (1950 = 100)						
Japan	100	170 2	253 9	70 2	49 2	153 9
U S	100	136 2	177 3	36 2	30 2	77 3
Percentage employed in agriculture						
Japan	48 7	30 0	17 8	−38 4	−40 7	−63 4
U S	12 1	8 3	4 6	−31 4	−44 6	−62 0
Percentage female in non-agricultural employment						
Japan	29 0	35 1	36 6	21 0	4 3	26 2
U S	31 3	34 6	38 3	10 5	10 7	22 4

SOURCES —U S data U S Bureau of the Census *Statistical Abstract of the United States* (various issues and U S Department of Health, Education, and Welfare (1968*a*), Japanese data Japan Prime Minister s Office *Statistical Yearbook of Japan* (various annual issues) and Mainichi Newspaper Company (1970

* Average between 1965 and 1970
† Average between 1949 and 1951
‡ Average between 1957 and 1960
§ Average between 1965 and 1968
‖ For 1968

decline in terms of price and income variables, but they are not likely to facilitate understanding of the shorter-term movements by which the decline may actually occur

In table 10 some of the salient features of the United States and Japanese experiences during the postwar years are compared If the economic framework for fertility analysis has general validity, it would predict more rapid fertility changes for countries experiencing more rapid changes in their economies Clearly, both the economic variables examined and fertility rates changed more rapidly in Japan than in the United States during 1950–69 A similar point is observed if one confines the comparison to the periods in which fertility declined 1950–60 for Japan and 1960–69 for the United States.

Much more work is needed before we fully understand the reasons for the possible acceleration in the decline of fertility demand in Japan in the 1950s and the leveling off in the 1960s A promising area for investigation is the interaction of the demands for quality and quantity of children in Japan. There may have been an accelerated substitution of quality for quantity during the postwar years because of income effects

Also, the prospect of future economic growth may have raised the expected return to education and rising family income increasingly facilitated the financing of children's education beyond the level required by law. It is noteworthy that the enrollment rate in kindergarten (which is not compulsory) increased more sharply between 1950 and 1960 (by 95 percent per 5 years) than between 1960 and 1965 (by 58 percent) Also, the enrollment rate in senior high school (not compulsory) rose more rapidly between 1950 and 1960 (by 26 percent per 5 years) than between 1960 and 1965 (by 20 percent) The model may also be extended to include such interrelated aspects of behavior as marriage and female labor-force participation Such a model may be estimated using more disaggregated data, say by age, than I use here The rapid decline in Japanese fertility from 1950 to 1960 coincided with an increase in the labor-force participation of females and a decrease in the proportion of married women among women of childbearing age For example, among females, ages 20 to 39, living in urban areas, the labor-force participation rate increased from 38 percent in 1950 to 43 percent in 1960, and the proportion ever married decreased from 75 to 70 percent The most dramatic changes occurred for females, ages 20 to 24, among whom the proportion ever married dropped from 43 percent in 1950 to 31 percent in 1960 but showed no appreciable change in the 1960s The pace of all these developments appears synchronous with the fertility history of postwar Japan, and possible relationships among them deserve rigorous analysis

Finally, it would greatly enhance the scope and accuracy of analysis to have micro, or household, data for Japan on such related characteristics as fertility, wage rates and family wealth, age, and work experience, parents' and children's education, and contraceptive use Ideally, such data would be collected longitudinally for a given set of families and made available to interested researchers in raw rather than summarized form

Comment

Gary R. Saxonhouse

University of Michigan

At a conference of distinguished specialists on human capital theory my comparative (absolute?) advantage is in the study of the Japanese economy. Much the same can be said for Hashimoto's paper. No attempts at any significant new departures in human capital theory or in the testing of such theory are made in this paper. Rather, one finds here an extension of a now-familiar approach to Japanese data with now-familiar results.[1] This is a study that can stand alongside the Israel and Taiwan studies by Ben-Porath (1973) and T. Paul Schultz (1973), respectively. Hashimoto's paper is another line, and an admirably executed line at that, in T. Paul Schultz's table A1 (1973).

Hashimoto's ultimate concern is in explaining the acceleration in the 1950s of the long-term decline in Japanese fertility and the subsequent bottoming out of this decline in the 1960s. The achievements of this particular paper are more modest. An attempt is made to challenge what Hashimoto finds is a simplistic overemphasis on the diffusion of abortion and contraceptive practice after 1948 (i.e., a downward shift in the supply of birth-control schedules). As an alternative, Hashimoto stresses outward shifts of the demand schedule for birth control resulting from the increasing cost of female time in postwar Japan.[2]

---

[1] Hashimoto states that, previous to this paper, there had been but one other economic study of fertility in Japan. This is a bit unfair. Only if the literature is restricted to that in English and one assumes the economic study of fertility starts with Becker is this statement true. Unlike the United States, in Japan throughout the twentieth century demographic study remained within the province of economists. For a sample of Japanese work, see Tachi and Arao (1935) and Tachi (1946).

[2] My reading of the Japanese literature suggests that most Japanese demographers would applaud Hashimoto's stress on changes in both the demand and supply schedules. Some Japanese demographers do appear to part company with Hashimoto when it comes to a discussion of the determinants of the demand-schedule shifts. There is a Japanese emphasis on the role of wealth. Relative to the period prior to World War II, Japan of the early 1950s is seen as a poor country.

I am much in sympathy with Hashimoto's hypotheses. There is considerable evidence to suggest that control of fertility within marriage was widespread long before 1948, and I do find stress on the cost of female time provocative.[3] At the same time, I do not think Hashimoto's various analyses are altogether persuasive. For example, if instead of using maximum crude birth rates and maximum general-fertility rates, maximum total-fertility rates for 1950 and 1960 are calculated, it appears that the practice of abortion might very well have tripled the rate of fertility decline.[4] The decline in the estimated maximum total-fertility rate between 1950 and 1960 is 16.8 percent. This is only a modest acceleration of the 12.6 percent rate of decline exhibited in the actual fertility rates between 1930 and 1940. An acceleration of this magnitude might easily be explained by the diffusion of modern contraceptive techniques during the 1950s. This mild acceleration might also be viewed as a reaction to the much higher-than-trend total-fertility rates between 1940 and 1950. Thus, it would appear that if Hashimoto wishes to stress the role of demand factors in the acceleration of Japanese fertility decline, he more than likely cannot avoid an inquiry into the extent to which legal abortions were substitutes for older forms of birth control.

Regardless of the magnitude of the acceleration remaining to be explained by some factor other than the legalization of abortion, I am not satisfied with Hashimoto's analysis of the spread of contraception. That the share of abortions in total births is high and rising in the rural regions of Japan in the presence of a sharply declining pregnancy rate can be taken as evidence that the spread of contraception is not sufficient to explain fertility changes unexplained by the legalization of abortion only if (1) abortion as a technique of birth control is already completely diffused, and (2) abortion is viewed as a technique of last resort.[5]

I doubt either of the above conditions held for Japan during the 1950s. Abortion diffused throughout Japan in much the same manner and only slightly in advance of modern contraception. At any point during that decade, for any age cohort, the better educated a couple and the larger the city in which a couple lived, the higher the probability they had experienced an abortion. Among couples who had never practiced contraception, the probability of an abortion experience was relatively low. Significantly, high rates of abortion were found among couples who had practiced contraception and for one reason or another discontinued this practice.[6] I would conclude that abortion was not necessarily a birth-control technique of last resort and that the rising share of abortions in total pregnancies in rural areas during the 1950s might very well reflect a positive reaction to the diffusion of a new technique.[7]

---

[3] For a discussion of the history of birth control in Japan, see Honjo (1941).

[4] The maximum total fertility rates in 1950 and 1960 are 4.16 and 3.46. I have calculated these rates with data taken from Japan Ministry of Health and Welfare (1967) and Aoki (1967).

[5] The industrial-rural dichotomy presented in Hashimoto's table 4 is based on a comparison of six urban prefectures with five rural ones. As an alternative, the share of induced abortions among pregnancies of Japanese women not living in cities of 50,000 or more (the rate typically regarded as the rural rate) can be examined. This rate rose throughout the 1950s, reaching a peak in 1961 (see Aoki [1967, p. 44]).

While Hashimoto does not make manifest the need for a new demand variable to explain the movement of postwar Japanese fertility, the final section of his paper does stand on its own as a test of competing hypotheses in explaining regional differentials. The results reported in tables 8 and 9 tend to support Hashimoto's de-emphasis of the role of the cost of fertility control. In both sets of regressions, I have a particular problem with the education variables Hashimoto chooses to use as explanatory variables. The male education variables in the completed-fertility regressions are based on the assumption that Japanese women are marrying Japanese men of the same age. Actually, Japanese women of the age cohorts in question were marrying men who were, on the average, 4 to 5 years older than themselves.[8] The peculiarities of the Japanese enforcement of their compulsory-education legislation make it a matter of some importance whether a male reached school age before 1900 (as a substantial proportion of the husbands of the women in these cohorts did) or after 1905. There was considerable variability among prefectures in the degree of male compliance with the compulsory education laws before 1900. After 1905, compliance was almost complete among the prefectures.[9] The degree of compliance across prefectures is only weakly correlated between the two periods. Interestingly enough, noncompliance for males in the earlier period is rather strongly correlated with noncompliance for females in the later period.[10] It follows that if one is a believer in the cost-of-fertility hypothesis (and only if one is a believer in the cost-of-fertility hypothesis), it is possible to argue that the errors in the male education variable are imparting a positive bias to the estimate of the coefficient of the no-education variable for females and a negative bias to the estimate of the coefficient of the no-education variable for males. These biases could be quite substantial and might account for Hashimoto's interesting results.

---

[6] The surveys on which the above analysis is based were conducted and periodically published by the Mainichi Newspaper Company (1950), see also Taeuber (1958, p. 282).

[7] There is every reason to believe that the spread of contraceptive methods in rural Japan was rapid in the late 1950s. The aforementioned Mainichi surveys indicate a better-than-25-percent increase in contraceptive practice between 1955 and 1959. These same years saw family-planning expenditures by Koseisho (Ministry of Health and Welfare) at their real height. Also, survey evidence analyzed by Aoki (1967, p. 57) suggests that increases in contraceptive efficiency were also greatest during this period.

[8] Summary statistics on the average age at first marriage, by sexes, are available in the annual reports, *Descriptive Materials on Japanese Vital Statistics*, published by the Japan Prime Minister's Office (1917*f*).

[9] In 1900, the percentage of the male school-age population attending primary school or otherwise complying with the compulsory-education laws ranged across prefectures from 79.1 to 97.5 percent. By 1905 the degree of compliance varied from 94.7 to 99.5 percent. These statistics are from annual reports published by Japan Ministry of Education (1904*f*).

[10] Using the 1960 census data, the simple correlation coefficient between the prefectural pattern of males with no education, age cohorts 50–59 and 60–69, is .39. The simple correlation coefficient between the prefectural pattern of males with no education in the age cohort 60–69 and females with no education in the age cohort 50–59 is .83.

Unlike the completed-fertility regressions, Hashimoto's regressions reported in table 9 deal entirely with postwar reproductive activities and specifically with 1960 live births. The education variables in these regressions refer to the entire (over age 15) prefectural population. Unfortunately, virtually the entire interprefectural variation in the proportion in the no-education variables relates to the experience of age cohorts that in 1960 were well beyond their reproductive years! I do not doubt that, inspired by Leibowitz's paper in this volume, work can be done on how a mother's education affects a daughter's fertility, but this is not the hypothesis in question.[11]

Also with respect to these same regressions, I am a bit puzzled by the use of the logarithm of earnings rather than arithmetic earnings. I suspect one gets a higher *t*-statistic on female wages when this form is used, but when using what should be thought of as grouped data, I am unsure of the theoretical justification.

In concluding my remarks, I would like to offer a suggestion for future research. Notwithstanding Hashimoto's very useful analysis, I do not find the postwar period to be the most interesting stretch of Japanese demographic history. While there are many instances of fertility decline in the context of rapid economic growth, the experience of Japan between 1875 and 1920 appears to be unique. There was rapid economic growth starting from a relatively high income base, a declining agricultural labor force, a revolution in female education coexisting with a high and, as recent research indicates, very likely a rising level of fertility.[12] Further, in view of the Japanese experience after 1920, I would submit that this episode constitutes a fascinating and important puzzle for those who would stress the role of economic factors in general, and the cost of time, in particular, in the explanation of fertility history

---

[11] Nationwide, 82 percent of men with no education and 96 percent of women with no education came from the age groups 50 and older and 40 and older, respectively. The simple correlation by prefecture for never-educated men above age 50 and younger than age 50 is .25. The simple correlation by prefecture for never-educated women above and below age 40 is .21.

[12] The official statistics collected for Japan suggest rising fertility throughout most of this period. These statistics have been traditionally treated with considerable skepticism on account of underreporting in the earliest periods. Taeuber, in analyzing these statistics, suggests that fertility probably rose substantially during the first part of this period and subsequently declined slowly (see Taeuber [1958, pp. 52–54]). Estimates which allow for the traditional criticisms of the official statistics have been prepared by Okazaki (1965). Okazaki estimates that the Japanese population in 1875 was almost 2 million above the official statistics. More recently, Akasaka Keiko has provided another series of alternative estimates. Somewhat surprisingly, in view of the earlier criticisms, she finds that the official statistics overestimate population by 600,000 in 1875 and understate the rate of population growth at least through 1910. For Akasaka's estimates, see Hitotsubashi University Institute of Economic Research (1968). For an informative discussion of the alternative estimates, see Umemura (1969).

Household and Economy: Toward a New Theory of Population and Economic Growth

---

Marc Nerlove

University of Chicago

> It is somewhat unusual to begin the treatment of a subject with a warning against attaching too much importance to it, but in the case of economics, such an injunction is quite as much needed as explanation and emphasis of the importance it really has It is characteristic of the age in which we live to think too much in terms of economics, to see things too predominantly in their economic aspect There is no more important prerequisite to clear thinking in regard to economics itself than is recognition of its limited place among human interests at large. [FRANK H KNIGHT, 1933 (1965)]

1. Introduction: Malthus Revised

Malthus, in essays published in 1798 and 1830, and the classical economists combined a very simple model of family decision making—procreation without bound except possibly by "a foresight of the difficulties attending the rearing of a family and the actual distresses of some of the lower classes, by which they are disabled from giving the proper food and attention to their children" (Malthus 1970, p 89)—with an equally simple model of the operation of the economy. According to the latter, a high level of capital accumulation induced by a high level of profits—representing the difference between output and the rent of land (natural resources) and wages—permitted a continual increase in output and population, albeit at the cost of resort to land of increasingly

---

I am indebted to Gary S Becker, Glen Cain, Richard B Freeman, Margaret G Reid, and T. W Schultz for discussions concerning the subject of this paper and for comments on a related paper. Virginia Thurner contributed valuable editorial advice. This research was supported by a grant from the Rockefeller Foundation to the University of Chicago for the study of the economics of population and family decision making

poorer quality, it did not, as the result of the model of family decision-making, lead to a rising standard of living for most people Thus, the classical economists achieved a very simple model of economic growth and development (Baumol 1970, pp 13–21) Modern growth theorists in the tradition of Solow (1956) and Swan (1956) have developed theories of economic growth based on far more elaborate theories of the economy, but few theories of population growth and family decision making have gone much beyond the Malthusian model (Pitchford 1973, pp 1–10) Although natural-resource constraints may be readily incorporated through the device of diminishing returns to scale in the variable factors (Swan 1956, pp 340–42), it is a constant proportional rate of population growth, perhaps aided and abetted by exogenous technological progress, that essentially drives the mechanism Discussions of the optimal rates of population growth or level of population do often attempt to integrate an endogenously determined population in the model But none, to my knowledge, have examined the feedback from changes in the economy and changes in the relative prices and costs which families face when they decide how many children they will have and what they will invest in those children's health, nutrition, or education, although Pitchford (1973) does discuss the costs of and returns to population control at the macro level

In recent years, the recognition, crucial to the understanding of long-term growth, that much investment which occurs in the economy is made in human beings rather than in physical capital and that fertility itself is shaped in important ways by economic considerations has led to renewed interest in the economics of the household decisions In that type of unit, not only decisions about fertility, but also those related to investments in human capital, consumption and savings, migration, labor-force participation, and, in a sense, marriage itself, are made

With this in mind, in Section 2 I briefly and critically describe the fundamental elements of the largely static theory of household production and choice This theory was developed in its modern form by Gary Becker (1965) and others, but most of its essentials originated in the much earlier work of Margaret Reid (1934), and it owes a good deal to Wesley Mitchell's seminal observations in his essay (1912) on "The Backward Art of Spending Money " Section 3 concludes this paper with some speculations on how the "new home economics" may be integrated in a theory of economic growth and development through an understanding of the way in which investment in human capital increases the value of human time and thus changes over time the resource constraints and the relative costs and prices which "households" face in their decisions on the number and quality of children they attempt to produce The question as to whether this constitutes a true economic explanation of the so-called demographic transition, and thus a revision of the Malthusian tradition, is basically an empirical one and is left open

2. The "New Home Economics": Summary and Critique

In its most unadorned form, economics is the theory of allocation of limited resources among competing ends in order to maximize satisfactions (or utility), subject to the constraints imposed by limitations in the availability of the resources required to achieve those ends. Various elaborations and accretions are necessary to accommodate this central theoretical core to the dynamics of choices made sequentially over time and in the presence of uncertainty regarding future constraints and future preferences At several points I shall have more to say about our present failure to reach successful accommodations in our underlying theory in these directions, as well as about our even more important limitations as economists to cope with the whole complex of issues raised by intergenerational transfers, within society as an ongoing concern, and particularly within the family

The first element in the new home economics is the utility function to be maximized Its form and its arguments (i e , what variables determine its level) are obviously crucial in determining the choices which result from its maximization But whose utility function is it that is maximized in connection with choices pertinent to marriage, children, consumption of commodities, work and leisure, and investment in all forms of capital? Considering the household as already formed, much of the theoretical underpinning rests on what I have called elsewhere the "Chicago model" (Nerlove 1972*b*) [1]

The "Chicago utility function," if I may call it that, has several key characteristics First, it does not involve nonmarket goods or physical commodities and purchasable services as we usually think of them in economics, but its arguments are abstract goods composed of a number of "attributes" which must themselves be produced within the household (Becker 1965, Lancaster 1966, Muth 1966) The importance of this characteristic of the utility function is that it leads directly to the key questions of household technology and the composition of different types of market goods and services and physical commodities, in terms of attributes contributing to satisfactions This, indeed, is the point made forcefully by Muth (1966) and recently emphasized by Michael and Becker (1972) Earlier (1947), Leontief pointed out that the theory of consumer behavior as then developed, although of great generality, lacked content to the extent that it gave no clue to the types of relations to be expected among different categories of goods

> The assumption of the existence of general categories of needs, different from demands for particular individual commodities,

---

[1] At this conference, Jacob Mincer objected to this and said that the model should really be called the "Morningside Heights" model, however, force of habit leads me to persist in this terminology.

> but still specific enough to be clearly distinguishable from each other, is basic to the man-in-the-street idea of consumers' demand One speaks of the desire for food as existing behind and separately from the particular demand for bread, apples, or Lobster à la Newburg This need for food is at the same time spoken of as something clearly distinguishable from the similarly general needs for clothing or, say, for shelter, each of the latter also thought of as existing separately although manifested through the particular demands for one-family houses, apartment flats, or woolen suits and raincoats [Leontief 1947, p 371]

Strotz (1957), in his introduction of the notion of a utility tree and in his later (1959) discussion with Gorman, attempted to give more empirical content to the theory of consumer demand in precisely this way, although he interpreted his results in terms of a "budgeting" process An important aspect of the household production model, including time and market-purchasable goods, as introduced by Becker (1965), Lancaster (1966), and Muth (1966), is precisely, as Muth points out, that it does yield "a utility function which is weakly separable when viewed as a function of commodities purchased on the market" (Muth 1966, p 700) While it is true, in a sense, that many conclusions of the "new home economics" could be derived directly from Strotzian utility trees and other specializing assumptions, the home-production aspect of the Chicago function lends an intuitive insight and empirical content which are lacking in the more abstract formulation It suggests a more direct look at the technology of processes within the household and particularly at how such processes use household time and nonpurchased inputs in addition to market-purchasable commodities Indeed, the supposed differences in the time intensity of the production of household goods give much of the content to recent applications of the "new home economics" to the problems of fertility and human capital formation [2] I return to this point in Section 3 where I speculate on how the increasing value of human time works through a Chicago utility function/household-production model to alter the behavior of generations through time

A second key characteristic of the Chicago utility function is that it is just that, one utility function—the welfare of the children and other members of the family is assumed to enter the utility function of a single decision maker (not always the husband and father[1]), thus obviating the assumption of a "family utility function" with all the concomitant problems of social utility functions in general. It is perhaps not entirely accurate to identify this position too closely with Chicago; it should perhaps be called the "Samuelsonian finesse" Samuelson (1956) writes.

---

[2] Unfortunately, to date relatively little attention has been paid to the implications of the household production model for the more general composition of consumption Time and market-purchasable-commodity intensities do differ greatly for different types of consumption (e g , drinking beer or going to a concert), and the new home economics has implications going beyond fertility and human capital formation (see, however, Michael [1972]).

> Where the family is concerned the phenomenon of altruism inevitably raises its head: if we can speak at all of the indifference curves of any one member, we must admit that his tastes and marginal rates of contribution are contaminated by the goods that other members consume These . . external consumption effects are the essence of family life . Such problems of home economics are, abstractly conceived, exactly of the same logical character as the general problem of government and social welfare [P 9]
>
> if within the family there can be assumed to take place an optimal reallocation of income so as to keep each member's dollar expenditure of equal ethical worth, then there can be derived for the whole family a set of well-behaved indifference contours relating the totals of what it consumes the family can be said to *act as if* it maximizes such a group preference function. [P. 21]

The problem with the Samuelsonian finesse, however, is that it assumes a fixed family membership, and a great deal of what the Chicago utility function is designed to explain is how that family composition gets determined This requires much more than Samuelson allows for in his formulation When, for example, are children members of the family, and thus codeterminers of the utility function, and when are they just arguments in the utility function determined for the family not including them? The full internalization argued by T W. Schultz (1972*b*) seems a necessary addition to the argument Yet, for this to be true, what might be called the "John Donne effect" must be extremely powerful Casual observation suggests that each individual's concern for others diminishes with distance in both time and space Yet it may be true under certain restrictive assumptions, as pointed out to me by Assaf Razin, that what might be called pairwise intergenerational internalization (by which I mean full internalization of the utilities of the next succeeding generation by the immediately preceding generation) would lead to essentially the same type of problem as that encountered in the discussions of optimal growth with an infinite horizon.

Morishima (1970, pp. 213–25) presents an extended discussion of some of the more technical issues involved in formulating dynamic utility functions and the conditions under which such functions can be reduced to the sum of discounted utilities of each future generation or at each future point in time, irrespective of the generation involved. In general, [illegible] conditions are highly restrictive and closely related to the conditions of Strotz (1959) and Gorman (1959) for strong separability of the utility function. But as Koopmans (1967, p 96) points out, the problem is really in some sense an ethical one

> What is at issue is clearly an intertemporal distribution problem that of balancing the consumption levels of successive generations, and of successive stages in the life-cycle of a given cohort of contemporaries The most pertinent decisions—individual, corporate, or governmental—are those that determine investment in physical capital, in human capital, and in research and development Investments in physical capital, if well made, augment future consumption through an increase in future capital-labor ratios Investment in human capital raises the quality of labor and, one hopes, of life Successful research and development augment future capital and labor inputs through the development of better techniques of production

Even regarded as a strictly behavioral model, pairwise internalization, which seems central to the Chicago utility function, has most important implications for the intergenerational transfers of material wealth and human capital which Knight (1921, pp 374–75) has so eloquently described as central to the continuity of the social order [3] Some of these implications bear on the issues discussed in the next section Here, however, it seems essential to point out the profound problem in, on the one hand, internalizing all the family members' satisfactions in one utility function and, at the same time, using this same utility function to determine the number and "quality" of the family members themselves Essentially the problem results from the condensation of a sequential, dynamic set of decisions into a theory of choice based on the maximization of a single, static, timeless utility function In its most extreme form, the issue of the conceptual adequacy of the approach arises in connection with the application of the new home economics to household formation—a formation which in an earlier, less aberrant, and nonconformist era might have been described as "marriage"[4] Yet marriage, in some sense, remains very much associated with procreation, and the act of marriage, or at least of household formation, is the normal first step in the central process of choice with which the new home economics deals Where then is the utility function? Can the entire process really be separated into two distinct parts, what the econometrician would call a recursive system? If a fundamental purpose of marriage is the procreation of a couple's own legitimate children, given a society's definition of both marriage and legitimacy, clearly the process cannot be regarded as recursive.[4] Again, the static character of the analysis, while not necessarily limiting its usefulness in an empirical sense, introduces a conceptual difficulty of a high order.

---

[3] Indeed, there is good reason to suppose that the reason for many institutions of society is precisely to ensure that the interests of future generations will be adequately guarded by the present. A meaningful theory of intergenerational welfare comparisons has not yet been developed, but such a theory is surely central to our understanding of household decision making and its consequences

The second element in the new home economics is the technology of household production described by a production function or functions and a list of the resources utilized in the processes involved. Typically, following Becker (1965), the inputs are time, perhaps distinguishable by household member (e.g., husband and wife) and market-purchasable commodities. These inputs are used to produce within the household the goods and services that in turn lead to satisfaction. In the simplest form in the economic theory of fertility, two time inputs (the husband's and wife's) and one general market-purchasable commodity are assumed in the household technology to produce three household goods: child numbers, child quality, and a general commodity called "other satisfactions" (Willis 1973; DeTray 1973). In all analyses to date an important further simplification of this basic technological structure has been made: each good entering the utility function is assumed to be produced by a separate independent production process. Jointness in production arises not because of common overhead factors within the household, but because the factor inputs available to the household are subject to overall constraints.[5] Willis (1973) and DeTray (1973), for example, both assume that child quality per child and child numbers are produced by independent production processes. Each factor separately enters the household utility function in the Willis formulation. The one is simply multiplied with the other to arrive at the final good, child services, which is assumed to enter the household utility function in DeTray's formulation. Both child numbers and child quality are generally assumed to be mother's time-intensive.

This formulation also neglects the sequential and essentially dynamic character of investments in child quality. Casual observation suggests that in many societies and families, the eldest son is often the beneficiary of the bulk of the investment in the human capital that takes place in the second generation Indeed, Sloan's recent work (1971, pp 31–32) suggests that in poor countries the sex, as well as birth order of a child, may be an important determinant of the infant's survival probabilities, presumably because of the differential investments in health and nutrition that take place within the family rather than for purely genetic reasons

---

[4] Still, it may be helpful analytically to tackle the problem in a stepwise fashion (see Becker 1973). In part of his work, Becker does consider the division of the gains from marriage, the most important gains being the children, and this is, of course, a matter closely related to the utility function maximized within the marriage.

[5] A partial exception is the work of Michael Grossman (1971; 1972, pp. 74–79). In his analysis Grossman does assume separate production processes but stipulates that the entire amounts of certain factors available to the household must enter each production function as one of its arguments. That is to say, in the standard analysis, if one could separately identify the amount of one factor—say, wife's time—used in a particular activity, one could measure the effect of a unit increase in that input on the output of that activity independently of the overall resource constraint on the wife's time to the household, whereas in Grossman's formulation the overhead factor, say, the family house, cannot in principle be allocated among the activities but must enter fully into each one. Grossman's work, however, still does lack generality in the respect discussed above, since complementarity among final goods cannot be encompassed in his formulation except through the rather Ptolemaic device of introducing some of the final outputs as inputs in the production processes of some of the other outputs.

The restrictiveness of the rather special assumption of separability of productive activities has not been apparent in the basically two-good (child services/other satisfactions) static models usually considered, for what is ruled out is complementarity among different outputs in a multiproduct context As is well known, under conditions of variable proportions, complementarity (in the sense that the output of one good could be increased without decreasing the output of the other good and without using additional resources) cannot occur if the production unit, in this case the household, is at an optimum and using resources fully In a three-or-more-product case, however, it is possible for several of the outputs to be complementary with each other, although substitutable jointly against the rest (Hicks 1939, p 92) [6] In the latter case, holding available resources constant, increasing the level of one of the outputs optimally might well involve also increasing another, although clearly a third output level would have to be decreased One can easily see how restrictive the elimination of complementarity could be in a context involving several dimensions of child quality, for example, health and physical development and intellectual achievement In a dynamic context such potential complementarities are of even greater significance since there is, for example, some evidence that early underinvestment in nutritional capital may substantially affect the productivity of later investments in intellectual human capital (Berg 1973, pp 9–12) [7]

The third element in the new home economics is a set of assumptions about the way in which household resources, principally time, can be transformed into market-purchasable commodities to be used in the household production process Strictly speaking, I suppose one could consider this set of assumptions as part of the general technology of household production and subsumed under the second element of the theory It is, however, better to treat the matter separately, since most of what is involved concerns the terms upon which household members can enter the labor market, the wages they can earn, and, somewhat secondarily, the prices at which market commodities can be purchased It is here that the lack of dynamic character of the new home economics cuts most

---

[6] In Hick's argument, substitute the word "outputs" for "factors" and vice versa, and hold the factor of production (his "output") fixed

[7] It would seem more straightforward and perhaps more desirable in future work to deal with simple forms of multiproduct production functions For an empirical example, see Eads, Nerlove, and Raduchel (1969)

deeply into its potential implications for the central problems of fertility and female labor-force participation. The timing and spacing of children, the opportunities for part-time work and accumulation of lifetime labor market experience, and choices as to the amount of education to be invested in early in the life cycle all revolve heavily on the terms under which women can participate in the labor market and thus share in the transformation of the household's time resources into market commodities The human capital literature (Ben-Porath 1967; Mincer 1970, 1973) is, of course, replete with dynamic analyses of investment in human capital over time and the life-cycle effects of these investments on earnings But little of this work has entered the more general framework of the "new home economics," particularly as this theory bears on decisions concerning the numbers of children and their timing and spacing within a marriage and the relation of these decisions to the accumulation of other forms of assets The work of Heckman (1971), Ghez and Becker (1972), and unpublished work of Frank Stafford does, however, represent a notable beginning of the extension of this part of theory into more dynamically relevant realms [8] Yet we need to understand far more than we presently do about why the labor market functions so differently for men than for women, the role of institutional constraints, discrimination, and the relation of these to women's choices of occupation and timing of labor-force participation Once again, the simultaneities of the system severely limit our ability to break out a single segment for proper analysis

The fourth and final element in the new home economics is the resource constraints facing the household in its production and optimization decisions Traditionally these constraints are divided into time (husband's and wife's, although often the husband is assumed to devote full time to the market) and "other" nonwage income While it is universally recognized that some elements of household production and consumption—sleep and food, for example—are in fact inputs into the production-of-time resources, little attention has thus far been paid to the quality of the time resources and of other family resources—both genetic and material—passed from one generation to the next Arleen Leibowitz's study of Terman's 1921 sample of California school children, reported in this volume, bears importantly on the manner in which much human capital is passed from one generation to another, especially to a child of preschool age. The sample is very unrepresentative, but it is instructive, for the investments in her child of a mother's time and the quality of those investments, as measured by her education, are found to affect appreciably later measures of the child's ability and future earning capacity

---

[8] Much recent work, however, is dynamic only in the relatively trivial sense of involving maximization over a number of periods of time without uncertainty concerning the values of future values of exogenous constraints to future decisions This is the sense in which a dynamic programming problem can be turned into an ordinary programming problem of much larger dimensions, apart from the computational difficulties thus introduced, the chief defect of this approach is that it fails to lay bare the sequential nature of the decision-making process in the way, for example, in which the recursive solution to the general dynamic programming problem of Bellman does In the absence of uncertainty, however, nothing essential is lost by the straightforward multiperiod extension of the basically static framework When values of future constraints are uncertain, then it does become essential to understand and incorporate the sequential character of the decision process (see Nerlove 1972*a*). So far this has not been done in the literature of the new home economics.

The resource constraints facing the household, once it is formed, are, of course, a product of the household formation itself and thus connected with my earlier remarks on the inseparability of this complex of issues, from those of family choice and decision making. But, more important it is in this area that the complex issue of intergenerational transfers figures most prominently We live, after all,

> in a world where individuals are born naked, destitute, helpless, ignorant, and untrained, and must spend a third of their lives in acquiring the prerequisites of a free contractual existence The fundamental fact about society as a going concern is that it is made up of individuals who are born and die and give place to others, and the fundamental fact about modern civilization is that it is dependent upon the utilization of three great accumulating funds of inheritance from the past, material goods and appliances, knowledge and skill, and morale Besides the torch of life itself, the material wealth of the world, a technological system of vast and increasing intricacy and the habituations which fit men for social life must in some manner be carried forward to new individuals born devoid of all these things as older individuals pass out The existing order, with the institutions of the private family and private property (in self as well as goods), inheritance and bequest and parental responsibility, affords one way for securing more or less tolerable results in grappling with this problem [Knight 1921, pp 374–75]

So the apparently simple theoretical construct of a time budget plus other income constraint to the household conceals beneath its serene and mathematically differentiable exterior the central problem of the continuity of society itself[9]

---

[9] Closely connected to the matter of what is inherited from the past and transferred from generation to generation is the problem of how individual tastes and preferences are formed and how they may change over time I have not mentioned this issue lest I be excommunicated from the economics profession! It is virtually part of the definition of what an economist is that he takes tastes as given, and I sometimes suspect that many of us require all tastes to be identical and assume that all differences among individuals arise from differences in the resource constraints those individuals face Indeed, one of the consequences of the conjectures and speculations presented in Section 3 is that the tastes and preferences of at least different generations, if not of individuals within each generation, could remain constant while the number of children per family declined over time due to changes in the value of human time induced endogenously

By themselves the four main elements of the theoretical structure of the new home economics—(1) a utility function with arguments which are not physical commodities but home-produced bundles of attributes, (2) a household production technology, (3) an external labor-market environment providing the means for transforming household resources into market commodities; and (4) a set of household resource constraints—are incapable of yielding a series of well-defined implications about the main problems of household behavior with which we are concerned It is only a framework within which to think about these problems Many special additional assumptions, some of which have been mentioned, must be added to the framework to arrive at empirically refutable propositions Moreover, the nature of the required additional specifications is intimately related to the peculiarities of the particular bodies of data to which the new home economics has been applied. These data range all the way from aggregate time-series data covering long periods of time, to cross-sectional census data for both large and small geographic regions at a point in time and over time, to household and family data based on individual interviews, with and without collection of retrospective information

If to the key simplification involved in the assumption of separable independent productive processes within the household, one adds the assumptions that young children are highly intensive of the mother's time in comparison with other activities within the home and older children are less intensive, and that for institutional or biological reasons the comparative advantage of the male partner in the acquisition of market-purchasable commodities significantly exceeds that of the female partner, a number of interesting implications of the theory emerge which are tolerably well supported by the empirical evidence so far analyzed Setting aside for the moment the inadequacy of the observed market wage to measure the cost of a nonworking woman's time, the immediate implication of the theory is that a rise in cost of mother's time for the family will cause a substitution away from time-intensive goods such as children and toward those requiring more inputs of market-purchasable commodities Indeed, if we further assume momentarily that males are completely specialized in market activities, changes in their wages represent pure income effects for the family, and we do then observe for families with working mothers positive association of family size with income and negative association with female wage rates To the extent that education serves as a proxy for the relative costs of time which may measure some of these costs less imperfectly, or at least in a fashion different from market wages, similar differences are observed between the effects of male and female educational attainments (Mincer 1963, Nerlove and Schultz 1970; Willis 1973, DeTray 1973; Ben-Porath 1973)

The interpretation of many of the results is somewhat complicated by the effects of marriage on the association of the educational attainments of husbands and wives, and by the fact that wives and mothers for the most part specialize entirely in home activities during only varying fractions of their lifetimes Moreover, the effects of female education seem unexpectedly and puzzlingly nonlinear (Ben-Porath 1973) To illustrate the nature of the difficulty involved, consider the following simple analysis of the effects of assortative mating We know that men of higher-than-average levels of educational attainment tend to marry women of higher-than-average levels of educational attainment Suppose, however, that tastes for children differ among the population, women with a high taste for children and a low taste for market-related activities will tend in general to seek and to receive less formal schooling than women with the opposite preference Presumably a man's taste for children in a society where nearly all men specialize in market-related activities will not be reflected very greatly in his formal schooling Some effect of the amount of his formal schooling may occur through early marriage, but he will certainly seek a mate with like preferences Thus, men with a given level of schooling with a high taste for children will tend to marry women with less schooling than the average associated with the level these men have achieved If, as is common, a husband's educational attainment is associated primarily with permanent income effects within the household, and his wife's is associated primarily with the opportunity cost of time, one can see that the negative effect of the opportunity cost of time as measured this way on fertility will tend to be exaggerated, holding male educational attainment constant, since the difference between the two across couples partly reflects differences in tastes which are unobservable and not included in the statistical analysis (see also T W Schultz [1973*b*, p S8] for a related comment)

Turning the analysis around, we can ask what implications the new home economics has for female labor-force participation The new home economics predicts what is perhaps the obvious the composition of a woman's family is strongly associated with her labor-force participation Typically, the number of a family's children under the age of 18 and the age of the youngest child are both strong predictors of a woman's labor-force participation These facts about family composition have been interpreted by Cain (1966), for example, as measures of the opportunity value of a mother's time in the home Gronau (1973) attempted a detailed and sophisticated analysis of just this proposition, using a subsample of the 1/1000 sample from 1960 United States Census, moreover, he examined the interaction of educational attainment with family composition Others (Smith 1972, Leibowitz 1972) have documented that highly educated married women participate to a greater extent in the labor force and work more hours when they do work than married women with less schooling Married women as a group also tend to withdraw from the labor force when they have children; this is an implication of the new home economics, on the assumption that children are more female time-intensive than other commodities produced within the home, but the rates at which women with different educational attainments withdraw is not the same During the child-rearing years, more highly educated women reallocate more hours to household production than do women with less education. Ben-Porath's (1973) finding of a U-shaped relation between education and labor-force participation for Israeli women with young children strongly suggests differences in the effects of education at different levels on the relative efficiencies of home and market production and also interactions between female education and child quality. But all of this has been insufficiently explored within the presently existing framework of the new home economics As Sweet (1968), Nerlove and Schultz (1970), and Hall (1973) have emphasized, educational investments, labor-force participation, and fertility must be viewed, at least partly, as simultaneously determined choices This is surely one of the most important implications of the new home economics and only partly negated by the latter's currently static character

Finally, if we regard, as I think we must, the grand problem of the new home economics as the explanation of the demographic transition, that is, the "economic and social processes and family behavior that accounts for the marked decline from very high birth and death rates to modern very low birth and death rates" (T. W Schultz 1973*b*, p S4), the new home economics does have some insights, albeit limited, to offer (O'Hara 1972) Clearly, a high probability of child mortality affects the costs of achieving a given family size, that is, the number of children surviving to a given age If it is assumed that on the whole parents achieve a greater (discounted) sum of satisfactions the longer a child survives and if declines in mortality result in greater relative increases in the conditional probabilities of survival from earlier ages to successively older ages, declines in mortality should tend, according to the new home economics, to generate a greater demand for children This need not lead to an increase in births, however, since such declines in mortality lower the cost of child quality relative to the cost of numbers of children Of course, the net effect must depend on the technology of the production of child numbers and child quality as well as on the relative importance of these in the utility function These factors, of course, may vary substantially from time to time, culture to culture, and place to place The elucidation of such effects, however, must surely constitute one of the central challenges to the empirical application of the new home economics To explore such effects fully, however, requires that the household decision-making process be accommodated in a model of economic growth and development, a subject to which we now turn

3. Household Decision Making and Economic Growth: Speculations and Conjectures

> "Would you tell me, please, which way I ought to go from here?"
>
> "That depends a good deal on where you want to get to," said the Cat [*Alice's Adventures in Wonderland*]

In his classic paper on "Diminishing Returns from Investment," Knight (1944) pointed out that "if new investment can be freely directed to all uses, i e., embodied in all types of productive agents indifferently, it will not be subject to diminishing returns" (p 33) Moreover, he stressed, as did Marshall before him, the concept of capital in human beings, and that "in the production of laborers the matter of 'quality' is far more important than that of quantity in the crude sense of numbers" (p 35) Although investment "freely directed" might not be subject to diminishing returns, certainly, under static circumstances, continued investment in any one particular direction ought eventually to result in a declining rate of return Yet, as T W Schultz (1973*b*) has emphasized, the rate of return appears to have diminished little, if at all, in response to a high level and even accelerated pace of investment in human capital, and, indeed, it may have actually risen (T W Schultz 1971, p 173)

We do not, of course, have any really accurate measure of the extent of investment in human capital as compared with investment in nonhuman capital and in the stock of knowledge through investment in research and development Yet, there are a number of clues which suggest that the capital stock invested in human beings, even on a per capita basis, has been a steadily growing portion of the total capital stock T W Schultz (1961, p. 73) suggests a rise of the value of the stock of educational human capital embodied in the stock of labor of persons age 14 and older from 18 percent in 1900 of the total educational and physical capital to 30 percent in 1957 These estimates do not include on-the-job training or investments in better health and nutrition In terms of gross capital formation, Kuznets (1966, p 243) calculates, on the basis of Schultz's earlier work, a rise in the share of investments in formal education alone from "about 9% in 1900 to over 38% in the 1950s" Moreover, for Western countries as a whole, Kuznets (1971) calculates the share of labor has risen from 55 to 75 percent of national income over the same period. These facts, meager as they may be, suggest two significant questions which are germane to the issue with which this paper began, namely, how can the new home economics be integrated into a general theory of economic growth and development in a manner which has some hope, at least, of bearing on the grand question of the demographic transition? These two questions are the following

First, what accounts for the failure of the rate of return to investments in human capital, even counting educational investments, to fall, despite a high and accelerating rate of investment in this form of capital relative to other forms? In other words, why does there appear to be a persistent disequilibrium among these rates of return?

Second, quite apart from the possibility of disequilibrium rates of return (even if they are in equilibrium), what effect will increasing human capital investment per capita have on the allocation of resources within the household and what, if any, repercussions will it have for the rate of growth of population and labor force? A number of speculations and conjectures on the answers to these questions follow, they emphasize the role of the increasing value of human time over time and its relation, which is reciprocal, to the increasing level of investment in human beings

Razin (1969, 1972) showed how, under certain circumstances, the ratio of human to total capital per capita would increase along the optimal growth path of an economy experiencing technical progress As indicated earlier, T W Schultz has emphasized this aspect of the persistant failure of rates of return to investments in human capital to decline. The demand for skills and knowledge embodied in human capital does not decline because of additional investments in the stock of useful knowledge and technique (technological change) which require the continual adaptation and adjustment of the human agent to utilize efficiently this augmented stock and seek out the new sources of investment opportunities which maintain the growth process But even in the absence of a persistent disequilibrium created by the demand for human capital, it is possible that the rates of return to such investments would fail to fall over time, or fall only slowly, in relation to the rates of return to other forms of investment because of endogenously changing relative cost of investment, that is, changes on the supply side

One of the most important consequences of the growing "quality" of human beings as reflected in the increased stock of human capital per capita, as pointed out by T W Schultz (1973*b*), is the increasing value of human time per unit of such time Many of the consequences of the increasing value of time over time are amusingly explored by Linder (1970) in his penetrating study of *The Harried Leisure Class.* Yet Linder and Schultz fail to note the important link which may exist between the increasing value of human time, due presumably to investment in human capital as well as to investment in other forms of capital and in technological change, and the terms on which investment in human capital takes place If one assumes, as I think plausible, that children (as regards both quality and quantity) are time-intensive as compared with other goods produced within the home, it follows that unless the increasing investment of human capital increases the marginal productivity of a unit of time in the care and rearing of children within the home in an offsetting fashion, increases in the value of time will lead to a shift away from children to less time-intensive activities To be sure, such a substitution effect may be offset by a strong income effect, but there are still further grounds to suppose that both substitution and income effects will tend to lead to an increase in child quality rather than child numbers

In a recent note (1973), Becker and Lewis explored the consequences of a simple model of the relation between the quality and quantity of a good (in this case, children) entering the utility function, under the assumption that quality per unit is the same for all units Increases in either quantity or quality cause the shadow prices of the other to rise and if, as is plausible, the income elasticity for quality is greater than that for quantity, the resulting increase in the shadow price of numbers would reduce the apparent income elasticity for quantity, perhaps even to negative level. Thus, the income effect of the increasing value of human time should, under these circumstances, lead to a substitution of quality for quantity of children, in addition to leading to a substitution away from children altogether to less time-intensive goods This rests of course, on the premise that the productivity of time in rearing children is unchanged or, at least, not greatly increased Moreover, the pure price effects are likely to work in the same direction, since it is plausible that quantity is more time-intensive than quality of children per unit of equal quality time and that, indeed, the shadow price of quality is likely to be reduced by increases in the quality of the mother's time input (I refer again to Arleen Leibowitz's paper in this volume)

The investments in child quality referred to earlier take two major forms (1) sound nutrition and health care, and (2) education, skills, or attitudes conducive to acquisition of further education and skills

Good nutrition and health care increase youngsters' chances of survival and may also affect their ability to absorb future investments in intellectual capital To the extent that such investments increase the life span, particularly the span of years over which a person can be economically active, such an increase in quality will raise the return to investments in human capital which sons and daughters may later wish to make in themselves To the extent that better health and nutrition result in a reduction in child mortality, they increase the satisfactions accruing to parents from other forms of investment which also raise child quality, for the returns to these investments may then be expected to be enjoyed over a longer period of time on average Increases in longevity, particularly of an individual's economically productive years, increase the amount of human time available without increasing population, such an increase would tend by itself to lower the value of time per unit, but, as we know, most of the effects of better health care and nutrition occur in childhood and enhance the quality of a unit of time in later years more than increasing the number of children On net balance, therefore, I would conjecture that better health and nutrition lower the costs of further investments in human capital relative to those in other forms of capital and increase the returns therefrom

The second main form which an increase in child quality may take is, as I have stated, through investment in the form of education, skills, or attitudes conducive to later acquisition of further education or skills. Much investment in human capital of this type tends to be time-intensive in the preschool years, although the productivity of a unit of a mother's time, as remarked, may be especially enhanced by a greater stock of human capital embodied in her, so it is not necessarily true that over time, as the result of the increasing value of human time, substitution will tend to occur away from this form of investment Nonetheless, it is in this area that we might expect some induced "technological innovations" which could economize on a mother's scarce time Nursery schools, day-care centers, and the proliferation of "educational" toys are perhaps examples On the whole, then, I think we may conclude that increases in quality of children relative to their numbers take the form of investments in human capital which ultimately have the effect of raising the value of time per unit in the economically active years of adulthood

For reasons which I feel certain we do not fully understand, but which are due in part to the presence of children's utilities in the utility function of the family to which they belong, parents do desire to bequeath a stock of capital to their children Since the stock of capital, material and intangible, human and nonhuman, is growing per capita in Western economies, one must assume that parents desire to pass along more than that which they received from their parents, or that institutions in the economy function in such a way as to induce this outcome Irrespective of the motivation, however, the increasing value of human time must have an effect on the form in which this capital is passed on As long as rates of return to investments in human capital remain above, or fall more slowly than, the rates of return to investments in other forms of capital, parents will be induced to bequeath a greater part in the form of human capital. Thus, the tendency toward increasing quality of children will be intensified by the bequest motive, despite the opposite tendency, resulting from the increasing cost of time, to invest in bequests which are less time-intensive But as rates of return tend to equality over time—if they ever do—parents should tend to bequeath less in the form of human capital and more in the form of financial and physical capital Nonetheless, as long as investment in human capital occurs, the value of a unit of human time will continue to rise with increases in the stock of capital per capita, reinforcing the tendency to fewer children of ever-higher quality Substitution will occur in favor of fewer children of higher quality and perhaps eventually against both quality and quantity of children in favor of commodities and knowledge The "facts" cited at the beginning of this section, suggesting an increasing portion of total capital formation in this century has occurred in the form of human capital, however, indicate that we may be far from the point at which such substitution begins to take place against children, quality, and quantity combined

The outlines of a revised Malthusian model begin to emerge, albeit dimly, from the foregoing conjectures and speculations In this model, the value of human time and changes in that value over time are pivotal, and the limitations imposed by natural resources are mitigated, if not eliminated, by technological progress and increases in the stock of knowledge and of capital, both human and nonhuman The main link between household and economy is the value of human time, the increased value of human time results in fewer children per household, with each child embodying greater investments in human capital which in turn result in lower mortality and greater productivity in the economically active years Such greater productivity in turn further raises both the value of a unit of time and income in the subsequent generation and enables persons of that generation to make efficient use of new knowledge and new physical capital Eventually, rates of return to investments in physical capital, new knowledge, and human capital may begin to equalize, but as long as investment occurs which increases the amount of human capital per individual, the value of a unit of human time must continue to increase It is not possible to say whether the diminishing ability of a human being to absorb such investment would eventually stabilize the number of children per household and at what level, given the satisfactions parents obtain from numbers of children as well as their quality Nonetheless, over time the model does predict in rough qualitative fashion declining rates of population growth (perhaps eventually zero rates or even negative rates for a time) and declining rates of infant mortality These are the main features of the demographic transition

Much remains to be done if this rough and speculative outline is to be translated into a true integration of the new home economics with the modern theory of economic growth, particularly as the former evolves along more dynamic and empirically relevant lines First, the model must be mathematized—and it is clear that there are many forms in which this may be accomplished—so that the crucial parameters and behavioral relations may be isolated These must then be studied empirically, for without quantitative knowledge of the parameters and technologies involved and the key behavioral relations the course of the important variables over time cannot be predicted or compared with past behavior

The role of the growth theorist is apparent in the first of these tasks and those of the econometrician and economic historian in the second Finally, the conceptual foundations of the new home economics, particularly with respect to intergenerational transfers, must be clarified if we are to understand the extent to which actual growth departs from the optimal path, the reasons for such departure, and whether collective action is either desirable or necessary to correct such departures.

I hope this paper represents at least a modest beginning in showing where it is we want to get to, if not a set of directions on which way we ought to go from here.

Comment

Zvi Griliches

Harvard University

I shall divide my comments into two parts, corresponding to Nerlove's discussion of the "new home economics" and the stylized "facts" that appear to be building blocks for a future growth model. I am very much in agreement with Nerlove's exposition of the "new home economics," but I would like to amplify a bit two of his implied criticisms of the state of this theory.

One of the major working assumptions of the theory is the existence of a common family utility function. In his paper in this volume, Gary Becker shows that by introducing the notion of "caring" (or the interdependence of utilities) one can show that the family will behave as if it has a common utility function. This will not do, I think, for the analysis of the empirical phenomena that we are really interested in. What parents care for is not the utility that their children receive, but the utility function that the children have and the resources that they control. Parents care about the consumption basket of their children, they have preferences over actual actions, not just their subjective outcomes. Much of the within-family conflict comes from different evaluations of the same consumption opportunities. For example, since many families appear to subsidize their children's higher education, and since the marginal valuation of family funds may not be the same for different family members, it may pay for the young both to pursue higher education further than appears warranted on straight rate-of-return calculations (somebody else is paying the cost), and to work less hard at it than the donors of the money would have liked them to do. In any case, a common utility function cannot explain either the growth of households or their dissolution, or indicate the point at which it pays for the young to opt out of it. It is probably also not necessary for the analysis of such questions, as is indicated by Becker's model of the marriage market.

The main shortcoming of the "new home economics" for the analysis of fertility decisions is that it assumes too little. The basic postulates are that children are goods, that all goods are subject to two constraints—time and money—and that children are relatively time-intensive goods. But this does not distinguish children from hi-fi sets! Adding the observation that children tend to use mother's time more than father's does not get us much further. Moreover, given the aims of the theory, this should probably come out as a conclusion, rather than be assumed from the beginning. In any case, the theory focuses on children as a good whose costs have a relatively high time component (particularly mother's time). Thus, any changes that have occurred have to be explained in terms of changes in total marginal cost (money and time) of this good (children) relative to the cost of all other consumption goods. I do not think that this is enough. The theory currently provides one explanation (among several competing ones) of the secular decline in fertility, attributing it to the rising cost of time, but it fails to provide any convincing explanation for the major fertility cycles that we have experienced. As the discussion at this conference made clear, we have no economic explanation yet within the framework of this theory either for the baby boom of the 1950s or the current rather sharp decline in fertility.

My belief is that if we want to study the demand for children, we have to put more content into the theory and start asking why do people want to have children, what are the returns and not just the costs of this activity? In terms of the theoretical framework used in this volume, I am looking for shifters of the utility function, or, alternatively, for factors that change the implicit household production function. If we are studying the demand for children rather than for hi-fi sets, we have to ask ourselves what it is about children that distinguishes them from other time-intensive durable goods. Perhaps we should go back to some low-level discussions about the "motives" for having children, along the lines of the discussion of the demand for money in older textbooks. I would distinguish at least three interdependent motives: (1) economic security (current labor and old-age provisions), (2) the production of reciprocal caring, and (3) an attempt at immortality via one's offspring.

The last motive explains the attempts to impose an image on the children (an improved version of our own) and the interest in the actual basket of consumption rather than just the summary state (utility) of their well-being. The returns in terms of the first two motives have been declining secularly, perhaps even faster than the rise in costs implied by the cost-of-time hypothesis. It is my guess that part of the recent sharp decline in fertility stems from the clear recognition on the part of the current childbearing cohort of the relatively low rate of return experienced by their own parents.

Turning to the last part of Nerlove's paper, I want to take issue with the two stylized facts that appear to be empirical building blocks for his growth model: the constancy of the rate of return to schooling and a persistent disequilibrium reflected in higher rates of return to the investment in human as compared with physical capital. Both Finis Welch and I have to take some blame for popularizing the first "fact" at a time when it was beginning to cease being a "fact." The persistence of relatively high rates of return to schooling through the 1950s and 1960s in the face of rising schooling levels throughout the economy was probably the result of an accidental constellation of forces rather than the expression of an underlying constancy in the economic mechanism. The higher education boom was sustained by three forces: (1) the first round of cohorts to be educated after World War II was a relatively small and declining fraction of the population, (2) the post–World War II baby boom increased greatly the demand for teachers at all levels, and (3) at about the same time the government superimposed on all of this a space-defense-research-and-development boom, heavily human capital-intensive, resulting in a scramble for young, educated talent. Unfortunately, these fortuitous influences have run their course. The educational system has probably reached its longer-run equilibrium level, if it has not overshot it. The space research and development boom is over now, at a time when there is, and will be for the next 5 years or so, an annual wave of an additional one million highly educated workers arriving at the doors of the full-time labor force. The rates of return to schooling have already started falling and will probably fall quite a bit further before supply response catches up with them.

Also, I am not sure that I understand the disequilibrium discussion in Nerlove's paper. First, I know of no study that shows that relevant rates of return, computed in comparable terms, are significantly higher for schooling than for physical investment. The fact that human capital may have been growing faster than physical is no evidence for this proposition, and it does not require it.

Combined References

Aigner, D "An Appropriate Econometric Framework for Estimating a Labor Supply Function from the SEO File " Madison Soc. Systems Res Inst , Univ Wisconsin, 1971

Alchian, Armen A , and Demsetz, Harold. "Production, Information Costs, and Economic Organization " *A E.R* 62 (December 1972) 777–95

Andersen, Ronald, and Anderson, Odin W. *A Decade of Health Services · Social Survey Trends in Use and Expenditures.* Chicago: Univ Chicago Press, 1967

Aoki, Hisao *Selected Statistics concerning Fertility Regulation in Japan* (In Japanese) Tokyo Ministry Health and Welfare, 1967.

——— "A General View of Fertility and Its Regulation in Japan " (In Japanese) *Jinko Mondai Kenkyu* 114 (April 1970): 5–20.

Aoki, Hisao, and Nakano, Eiko *Summary of the Fertility Surveys in 1940, 1952, 1957, and 1962* (In Japanese) Tokyo: Ministry Health and Welfare, 1967

Arrington, Leonard J. *Great Basin Kingdom* Lincoln. Univ. Nebraska Press, 1958

Ashenfelter, Orley A , and Mooney, Joseph D. "Graduate Education, Ability and Earnings " *Rev Econ and Statis* 50 (February 1968): 78–86.

Baumol, William J. *Economic Dynamics · An Introduction.* 3d ed New York: Macmillan, 1970

Becker, Gary S "An Economic Analysis of Fertility " In *Demographic and Economic Change in Developed Countries.* Universities–National Bureau Conference Series, no 11 Princeton, N.J . Princeton Univ. Press, 1960.

——— *Human Capital* New York Columbia Univ. Press (for Nat. Bur. Econ Res), 1964.

——— "A Theory of the Allocation of Time." *Econ. J.* 75 (September 1965): 493–517

———. *Human Capital and the Personal Distribution of Income: An Analytical*

Approach Woytinsky Lecture no 1 Ann Arbor Inst Public Admin , Univ Michigan, 1967.

——— "A Theory of Social Interactions " Unpublished paper Chicago Univ Chicago, September 1969

——— *Economic Theory* New York Knopf, 1971

——— "A Theory of Marriage Part I " *J P E* 81 (July/August 1973) 813–46

Becker, Gary S , and Chiswick, Barry R "Education and the Distribution of Earnings " *A E R* 56 (May 1966) 358–69

Becker, Gary S , and Lewis, H Gregg "On the Interaction between the Quantity and Quality of Children " *J P E* 81, no 2, suppl (March/April 1973) S279–88

Ben-Porath, Yoram "The Production of Human Capital and the Life Cycle of Earnings " *J P E* 75 (August 1967) 352–65

——— *Fertility in Israel, an Economist's Interpretation Differentials and Trends, 1950–1970* RM-5981-FF Santa Monica, Calif RAND Corp , August 1970 (*a*)

——— "On the Association between Fertility and Education " Mimeographed Research Report no 20 Jerusalem Hebrew Univ Jerusalem, 1970 (*b*)

——— "Economic Analysis of Fertility in Israel Point and Counterpoint " *J P E* 81, no 2, suppl (March/April 1973) S202–33

Berg, Alan *The Nutrition Factor. Its Role in National Development* Washington Brookings Inst , 1973

Blau, Peter M , and Duncan, Otis Dudley *The American Occupational Structure* New York Wiley, 1967

Bowen, W , and Finegan, T *The Economics of Labor Force Participation* Princeton, N J Princeton Univ Press, 1969

Bowles, Samuel "Schooling and Inequality from Generation to Generation " *J P E* 80, no 3, suppl (May/June 1972) S219–51

Box, G P , and Cox, D "An Analysis of Transformations " *J Royal Statis Soc* 26 (1964) 211–43

Box, G P , and Tidewell, P "Transformations of Independent Variables " *Technometrics* 4 (1962) 531–42

Broel-Plateris, Alexander "Marriage Disruption and Divorce Law " Ph D dissertation, Univ Chicago, 1961

Cain, Glen G *Married Women in the Labor Force An Economic Analysis* Chicago Univ Chicago Press, 1966

Cavalli-Sforza, L L , and Bodmer, W *The Genetics of Human Populations* San Francisco: Freeman, 1971

Cheung, Steven N S "The Enforcement of Property Rights in Children, and the Marriage Contract " *Econ J* 82 (June 1972) 641–57

Chiswick, Barry R *Income Inequality: Regional Analysis within a Human Capital Framework* New York· Nat Bur. Econ Res , 1973

Chiswick, Barry R , and Mincer, Jacob "Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985 " *J P E* 80, no. 3, suppl (May/June 1972) S34–66.

Chow, Gregory C "Tests of Equality between Sets of Coefficients in Two Linear Regressions " *Econometrica* 28 (July 1960) 591–605

Coale, A S , and McNeil, D R "The Distribution by Age of the Frequency of First Marriage in a Female Cohort." *J. American Statis Assoc* 67 (December 1972) 743.

DeTray, Dennis N *The Interaction between Parent Investment in Children and Family Size An Economic Analysis* R-1003-RF Santa Monica, Calif RAND Corp May 1972

——— "Child Quality and the Demand for Children " *J P E* 81, no 2, suppl (March/April 1973) S70–95

Dorjahn, V R "The Factor of Polygyny in African Demography " In *Continuity and Change in African Cultures,* edited by W. R Bascom and M J Huskovity Chicago Univ Chicago Press, 1959

Dugan, D J "The Impact of Parental and Educational Investment upon School Achievement " In *Proceedings of the Social Statistics Section, American Statistics Association, 1969* Washington American Statis Assoc , 1969

Duncan, Otis Dudley, Featherman, David L , and Duncan, Beverly *Socioeconomic Background and Achievement* New York· Seminar, 1972

Eads, George, Nerlove, Marc, and Raduchel, William "A Long-Run Cost Function for the Local Service Airline Industry. An Experiment in Nonlinear Estimation " *Rev Econ and Statis.* 51 (August 1969) 258–70

Ehrlich, Isaac "Participation in Illegitimate Activities· An Economic Analysis " Ph D dissertation, Columbia Univ , 1970

Fane, George "The Productive Value of Education in Agriculture in the U S " Ph D dissertation, Harvard Univ , 1972

Fisher, R A *The Genetical Theory of Natural Selection* 2d ed New York Dover, 1958

Freiden, Alan "A Model of Marriage and Fertility " Ph D dissertation, Univ Chicago, 1972

Fuchs, Victor. *Differentials in Hourly Earnings by Region and City Size, 1959* Occasional Paper no 101. New York Nat Bur Econ. Res., 1967

Gardner, Bruce. "Economic Aspects of the Fertility of Rural-Farm and Urban Women." *Southern Econ J.* 38 (April 1972): 518–24

Ghez, Gilbert R , and Becker, Gary S "The Allocation of Time and Goods over the Life Cycle." Report no 7,217 Chicago. Center Math. Studies Bus. and Econ., Univ. Chicago, April 1972.

Goldfield, S , and Quandt, R *Nonlinear Methods in Econometrics* Amsterdam North-Holland, 1972

Goode, William J *After Divorce* Glencoe, Ill Free Press, 1956

——— "Marital Satisfaction and Instability A Cross-cultural Class Analysis of Divorce Rates" *Internat Soc Sci J* 14, no 3 (1962) 507–26

——— *World Revolution and Family Patterns* New York Free Press, 1963

——— *Explorations in Social Theory* New York Oxford, 1973

——— *The Celebration of Heroes. Prestige as a Control System* (forthcoming)

Gorman, W M "Separability and Aggregation" *Econometrica* 27 (July 1959) 469–81

Griliches, Zvi, and Mason, William M "Education, Income, and Ability" *J P E* 80, no 3, suppl (May/June 1972) S74–103

Gronau, Reuben "The Intra-Family Allocation of Time The Value of the Housewives' Time" Research Report no 28 Jerusalem Hebrew Univ Jerusalem, June 1971

——— "The Wage Rates of Women—Selectivity Bias" Mimeographed New York Nat Bur Econ Res , 1972

——— "The Effect of Children on the Housewife's Value of Time" *J P E* 81, no 2, suppl (March/April 1973) S168–99

Grossman, Michael "The Economics of Joint Production in the Household" Report no 7,145 Chicago Center Math Studies Bus and Econ , Univ Chicago, September 1971

———. *The Demand for Health A Theoretical and Empirical Investigation* Occasional Paper no 119 New York Nat Bur Econ Res , 1972

Hall, Robert E "Comment" *J P E* 81, no 2, suppl (March/April 1973) S200–201

Hansen, W Lee, Weisbrod, Burton A , and Scanlon, William J "Schooling and Earnings of Low Achievers" *A E R* 60 (June 1970) 409–18

Harberger, Arnold "Taxation, Resource Allocation, and Welfare" In *The Role of Direct and Indirect Taxes in the Federal Revenue System* Princeton, N J Princeton Univ Press (for Brookings Inst and Nat Bur Econ Res), 1964

Hause, John "Earnings Profile Ability and Schooling" *J P E* 80, no 3, suppl (May/June 1972) S108–38

Heckman, James J "Three Essays on the Supply of Labor and the Demand for Market Goods" Ph D dissertation, Princeton Univ , 1971

——— "Shadow Prices, Market Wages, and Labor Supply" Paper presented at the Econometric Society Meetings, Toronto, December 1972. (Forthcoming in *Econometrica*)

Hicks, J R. *Value of Capital* Oxford Clarendon, 1939 (2d ed 1946)

Hill, C Russell, and Stafford, Frank P. "The Allocation of Time to Preschool Children and Educational Opportunity." Manuscript. Ann Arbor: Dept Econ , Univ Michigan, May 1973 (Forthcoming in *J Human Resources*)

Hitotsubashi University Institute of Economic Research. "National Income Study Group Materials C1, C18 " Mimeographed (In Japanese) Tokyo· Hitotsubashi Univ Inst Econ. Res , 1968

Honda, Tatsuo "Summary of the Third Fertility Survey in 1957 " (In Japanese) *Jinko Mondai Kenkyu* 77 (August 1959) 1–24

Honjo, Eijiro *Nihon jinko shi* [Japanese population history] Tokyo Nihon Hyoron-sha, 1941

Hotelling, Harold "Edgeworth's Taxation Paradox and the Nature of Demand and Supply Functions " *J P E.* 40 (October 1932) 577–616

Huffman, Wallace E "The Contribution of Education and Extension to Different Rates of Change " Ph D dissertation, Univ Chicago, 1972

Japan Ministry of Education *Yearbook of the Ministry of Education* (In Japanese) Tokyo. Ministry Educ , 1904 and succeeding annual issues

Japan Ministry of Health and Welfare *Vital Statistics of Japan* Tokyo Ministry Health and Welfare, 1970

Japan Ministry of Health and Welfare, Institute of Population Problems "Summary of the Fifth Mainichi Survey of Family Planning Opinion " (In Japanese) *Jinko Mondai Kenkyu* 77 (August 1959) 60–88

——— *Reproduction Indices for Japan, 1960–1965* (In Japanese) Tokyo: Ministry Health and Welfare, 1967

——— *Report on the Fifth Fertility Survey* (In Japanese) Tokyo Ministry Health and Welfare, 1968

Japan Prime Minister's Office *Descriptive Materials on Japanese Vital Statistics* (In Japanese) Tokyo Prime Minister's Office, 1917 and succeeding issues

——— *Annual Statistics of Japan 1960* Tokyo Prime Minister's Office, 1960 (*a*)

——— *Censuses of Population of Japan* Tokyo Prime Minister's Office, 1955, 1960 (*b*), 1965, 1970

——— *Statistical Yearbook of Japan* Tokyo Prime Minister's Office, various annual issues

Johnson, George E , and Stafford, Frank P "Social Returns to Quantity and Quality of Schooling " *J Human Resources* 8 (Spring 1973) · 139–55

Kagan, Jerome S "Inadequate Evidence and Illogical Conclusions " *Harvard Educ Rev* 39 (Spring 1969) 274–77

Kagan, Jerome S , and Moss, H A. "Parental Correlates of Child's IQ and Height." *Child Development* 30 (September 1959): 325–32.

Keeley, Michael C. "A Model of Marital Formation· The Determinants of the Optimal Age at First Marriage and Differences in Age at Mariage " Ph D dissertation, Univ. Chicago, 1973.

Kendall, M G "Proof of Relations Connected with the Tetrachoric Series and Its Generalizations " *Biometrica* 32 (1941). 196–99

Khaldi, Nabil "The Productive Value of Education in the U S Agriculture, 1964." Ph.D dissertation, Southern Methodist Univ , 1972

King, Willford Isbell *Employment Hours and Earnings in Prosperity and Depression—United States, 1920–1922* New York Nat Bur Econ Res , 1923.

Knight, Frank H *Risk, Uncertainty and Profit* Boston Houghton-Mifflin, 1921.

——— "Diminishing Returns from Investment " *J P E* 52 (March 1944) 26–47

——— *The Economic Organization* New York Harper & Row, 1965 (Originally published by the College of the University of Chicago, 1933)

Kogut, Edy L. "The Economic Analysis of Demographic Phenomena A Case Study of Brazil " Ph D dissertation, Univ Chicago, 1972

Koopmans, Tjalling C "Intertemporal Distribution and 'Optimal' Aggregate Economic Growth " In W Fellner et al , *Ten Economic Studies in the Tradition of Irving Fisher* New York Wiley, 1967

Kosters, M "Income and Substitution Effects in a Family Labor Supply Model " P-3339 Santa Monica, Calif RAND Corp , 1966

Kuznets, Simon *Modern Economic Growth Rate, Structure, and Spread* New Haven, Conn Yale Univ Press, 1966

——— *Economic Growth and Nations* Cambridge, Mass Harvard Univ Press, 1971

——— "Modern Economic Growth Findings and Reflections " (Nobel Prize lecture, December 1971) *A E R* 63 (June 1973) 247–58

Lancaster, Kelvin J "A New Approach to Consumer Theory," *J P E* 74 (April 1966) 132–57.

Leibowitz, Arleen "Women's Allocation of Time to Market and Nonmarket Activities. Differences by Education " Ph D dissertation, Columbia Univ , 1972

Leontief, W "Introduction to a Theory of the Internal Structure of Functional Relationships " *Econometrica* 15 (October 1947) 361–73

Lewis, H Gregg. "Income and Substitution Effects in Labor-Force Participation and Hours of Work " Mimeographed Chicago Dept Econ , Univ Chicago, 1972

Linder, Staffan B *The Harried Leisure Class* New York Columbia Univ Press, 1970

Low, S., and Spindler, A. *The Child Care Arrangements of Working Mothers* Children's Bureau publication no 461 Washington Government Printing Office, 1968

McCall, J. J "Economics of Information and Job Search " *Q J E* 84 (February 1970)· 113–26

Mainichi Newspaper Company. *Family Planning Survey* (In Japanese) Tokyo· Mainichi Newspaper Company, 1950 and succeeding years

——— *Nihon no Jinko Kakumei* [Japan's population revolution] (In Japanese) Tokyo Mainichi Newspaper Co , 1970

Malthus, T R *An Essay on the Principle of Population and a Summary View of the Principle of Population* Baltimore: Penguin, 1970 (Originally published in 1798 and 1830)

Marshall, Alfred *Principles of Economics.* 8th ed (1st ed 1890) London: Macmillan, 1930

Matsumoto, Y Scott, Koizumi, Akira, and Nohara, Tadahiro "Condom Use in Japan " *Studies in Family Planning* 3 (October 1972) 251–55

Michael, Robert T "Education and Fertility " Mimeographed New York. Nat Bur Econ Res , 1971

——— *The Effect of Education on Efficiency in Consumption* Occasional Paper no 116 New York Nat Bur Econ Res , 1972

——— "Education and the Derived Demand for Children." *J P E* 81, no 2, suppl (March/April 1973) S128–64

Michael, Robert T , and Becker, Gary S "On the New Theory of Consumer Behavior " Mimeographed New York Nat Bur Econ Res , December 1972

Michael, Robert T , with Edward P. Lazear "On the Shadow Price of Children " Mimeographed New York Nat Bur Econ Res , December 1971 (Presented at the 1971 meeting of the Econometric Society, New Orleans, December 1971)

Mincer, Jacob "Labor Force Participation of Married Women " In *Aspects of Labor Economics,* edited by H Gregg Lewis Universities–National Bureau Conference Series no 14 Princeton, N J Princeton Univ Press, 1962

——— "Market Prices, Opportunity Costs, and Income Effects " In *Measurement in Economics · Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld,* edited by Carl Christ et al Stanford, Calif . Stanford Univ Press, 1963

——— "The Distribution of Labor Incomes A Survey with Special Reference to the Human Capital Approach " *J Econ Literature* 8 (March 1970) 1–26

——— *Schooling, Experience, and Earnings* New York: Nat Bur Econ. Res , 1974

Mitchell, Wesley C. "The Backward Art of Spending Money " *A E R* 2 (June 1912): 269–81

Morishima, Michio *Theory of Economic Growth* 2d rev. impression Oxford: Clarendon, 1970

Mortensen, D T "Job Search, the Duration of Unemployment and the Phillips Curve " *A.E.R.* 60 (December 1970): 847–62.

Mosteller, Frederick *On Equality of Educational Opportunity* New York Vintage, 1972

Muramatsu, M "Effect of Induced Abortions on the Reduction of Birth in Japan " *Milbank Memorial Fund Q* 38 (April 1960) 153–66

——— "Problems in Procuring Contraceptive Materials in Rural Areas in Japan " In *Research in Family Planning*, edited by Clyde V Kiser Princeton, N J Princeton Univ Press, 1962

——— "Policy Measures and Social Changes for Fertility Decline in Japan " In *World Population Conference, 1965* Vol 2 New York United Nations, 1967

Muth, Richard F "Household Production and Consumer Demand Functions " *Econometrica* 34 (July 1966) 699–708

National Council of Churches, Bureau of Research and Survey *Churches and Church Membership in the U S* Ser A, no 3 New York Nat Council Churches, 1956

Nelson, P J "Information and Consumer Behavior " *J P E* 78 (March/April 1970) 311–29

Nelson, R , and Krashinsky, S "The Organization and Planning of Day Care for Children " Washington Urban Inst , August 1972

Nerlove, Marc "Lags in Economic Behavior " *Econometrica* 40 (March 1972) 221–52 (*a*)

——— "Tuition and the Costs of Higher Education· Prolegomena to a Conceptual Framework " *J P E* 80, no 3, suppl (May/June 1972) S178–218 (*b*)

Nerlove, Marc, and Schultz, T Paul *Love and Life between the Censuses A Model of Family Decision Making in Puerto Rico, 1950–1960* RM-6322-AID Santa Monica, Calif RAND Corp , September 1970

O'Hara, Donald J *Changes in Mortality Levels and Family Decisions regarding Children* R-914-RF Santa Monica, Calif RAND Corp , February 1972

Okazaki, Yoichi "Meiji Period Population " (In Japanese) *Keizai Kenkyu* 16 (July 1965) 207–13

——— "An Analysis of the Socio-economic Factors Affecting Fertility in Japan " (In Japanese) *Jinko Mondai Kenkyu* 114 (April 1970) 21–34

Parnes, Herbert S ; Shea, John R , Spitz, Ruth S , and Zeller, Frederick A *Dual Careers* Vol 1 Manpower Research Monograph no 21 Washington. Dept Labor, 1970

Pitchford, J. D *Population in Economic Growth* Mimeographed monograph Canberra· Australian Nat Univ , 1973

Polachek, Solomon "Work Experience and the Difference between Male and Female Wages " Ph D dissertation, Columbia Univ , 1973

Rahm, M "The Occupational Wage Structure " Ph D dissertation, Columbia Univ., 1971

Razin, Assaf. "Investment in Human Capital and Economic Growth. A Theoretical Study" Ph.D dissertation, Univ Chicago, 1969.

——— "Investment in Human Capital and Economic Growth" *Metroeconomica* 24 (May/August 1972): 101–16

Reid, Margaret G *Economics of Household Production.* New York: Wiley, 1934

——— *Housing and Income* Chicago Univ Chicago Press, 1962

Rheinstein, Max *Marriage Stability, Divorce, and the Law* Chicago Univ Chicago Press, 1972

Rivers, W H *The Todas* London Macmillan, 1906

Rogers, Daniel C "Private Rates of Return to Education in the U S · A Case Study" *Yale Econ Essays* 9 (Spring 1969) 89–134

Rosen, Sherwin "Short-Run Employment Variation on Class-I Railroads in the U S, 1947–63" *Econometrica* 36 (July/October 1968) 511–29

Ross, Sue "The Effect of Economic Variables on the Timing and Spacing of Births" Ph D research in progress, Columbia Univ, 1973

Ruderman, F *Child Care and Working Mothers* New York Child Welfare League America, 1968

Salzano, F M, Neel, J V, and Maybury-Lewis, D "Further Studies on the Xavante Indians" *American J Human Genetics* 19 (July 1967) 463–89

Samuelson, Paul A "Social Indifference Curves" *Q J E* 70 (February 1956) 1–22

Schoggen, P H, and Schoggen, Maxine F "Behavior Units in Observational Research" Paper presented at the symposium on Methodological Issues in Observational Research, American Psychological Association, San Francisco, 1968

Schultz, T Paul "Explanation of Birth Rate Changes over Space and Time A Study of Taiwan" *J P E* 81, no 2, suppl (March/April 1973) S238–74

Schultz, Theodore W "Education and Economic Growth" In *Social Forces Influencing American Education, 1961,* edited by Nelson B Henry Sixtieth yearbook of the National Society for the Study of Education, pt 2 Chicago Univ Chicago Press, 1961.

——— *Transforming Traditional Agriculture* New Haven, Conn Yale Univ. Press, 1964

——— *Investment in Human Capital* New York· Free Press, 1971

——— *Human Resources Fiftieth Anniversary Colloquium 6* New York· Nat Bur Econ Res, 1972 (*a*)

——— "The Increasing Economic Value of Human Time" *American J Agricultural Econ* 54 (December 1972): 843–50 (*b*)

——— "Explanation and Interpretations of the Increasing Value of Human Time" Mimeographed. Woody Thompson lecture, Midwest Economics Association, Chicago, April 5, 1973 (*a*)

——— "The Value of Children An Economic Perspective" *J P E* 81, pt 2, suppl (March/April 1973) S2-13 (*b*)

———, ed "New Economic Approaches to Fertility" *J P E* 81, pt 2, suppl (March/April 1973) (*c*)

Schwartz, Aba "Migration and Lifespan Earnings in the U S" Ph D dissertation, Univ Chicago, 1968

Shaw, George Bernard *Man and Superman* In *The Collected Works of Bernard Shaw* Vol 10, Ayot St Lawrence ed New York Wise, 1930

Shell, K *Essays on Theory of Optimal Economic Growth* Cambridge, Mass M I T Press, 1967

Shinozaki, Nobuo "Regional Problem of Population in Japan from the Viewpoint of Fertility" (In Japanese) *Jinko Mondai Kenkyu* 118 (April 1971) 5-20

Silver, Morris "Births, Marriages, and Business Cycles in the United States" *J P E* 73 (June 1965) 237-55

——— "Births, Marriages and Income Fluctuations in the United Kingdom and Japan" *Econ Development and Cultural Change* 14 (April 1966) 302-15

Sloan, F "Survival of Progeny in Developing Countries An Analysis of Evidence from Costa Rica, Mexico, East Pakistan, and Puerto Rico" R-773-AID Santa Monica, Calif RAND Corp, July 1971

Smith, James P "The Life Cycle Allocation of Time in a Family Context" Ph D dissertation, Univ Chicago, 1972

Solow, Robert M "A Contribution to the Theory of Economic Growth" *Q J E* 70 (February 1956) 65-94

Sontag, Lester W, Baker, Charles T, and Nelson, Virginia L *Mental Growth and Personality Development A Longitudinal Study* Monograph of the Society for Research in Child Development 23, no 68 Lafayette, Ind Child Development Publications, 1958

Stigler, George J "The Economics of Information" *J P E* 69 (June 1961) 213-25

Strotz, R "The Empirical Implications of a Utility Tree" *Econometrica* 25 (April 1957) 269-80

———. "The Utility Tree—Correction and Further Appraisal" *Econometrica* 27 (July 1959) 482-88

Suter, Larry E, and Miller, Herman P "Components of Income Differences between Men and Women in the United States" Paper presented at the American Sociological Association meetings, Denver, 1971

Swan, T W "Economic Growth and Capital Accumulation" *Econ Record* 32 (November 1956) 334-61

Sweet, James A "Family Composition and the Labor Force Activity of Married Women in the United States" Ph D dissertation, Univ Michigan, 1968.

Tachi, Minoru. "The Relationship between Production Increase and Population Growth among Japanese Prefectures, 1925–1935" (In Japanese) *Jinko Mondai Kenkyu* 5 (February 1946): 15–44

Tachi, Minoru, and Arao, M "Marriage Rate and Business Cycle in Japan" *Jinko Mondai* 2 (October 1935) 205–20

Taeuber, Irene *The Population of Japan* Princeton, N J . Princeton Univ. Press, 1958

Taubman, Paul, and Wales, Terence "Education as an Investment and Screening Device." Mimeographed New York Nat Bur Econ Res , 1972

Terman, Lewis M *Mental and Physical Traits of a Thousand Gifted Children* Genetic Studies of Genius, vol 1 Stanford, Calif : Stanford Univ Press, 1925

Terman, Lewis M , Burks, Barbara S , and Jensen, Dortha W *The Promise of Youth* Genetic Studies of Genius, vol 3 Stanford, Calif Stanford Univ Press, 1930

Terman, Lewis M , and Oden, Melita H *The Gifted Group at Mid-Life* Genetic Studies of Genius, vol 5 Stanford, Calif : Stanford Univ. Press, 1959

Umemura, Mataji "Meiji Period Population Growth" In *Keizaishi ni okeru jinko* [Population in economic history], edited by H Komatsu Tokyo Keio tsushin, 1969

U S , Bureau of the Census *County and City Data Books, 1962 Statistical Abstract,* suppl Washington Government Printing Office, 1962

——— *Statistical Abstract of the United States* Washington Government Printing Office, various issues

——— *U S Census of Housing, 1960* Vol 1 *States and Small Areas* Pt 1 *United States Summary* Washington Government Printing Office, 1963 (*a*)

——— *U S Census of Population 1960* Vol 1 *Characteristics of the Population,* pts 1–52 Washington Government Printing Office, 1963 (*b*)

——— *U S Censuses of Population and Housing, 1960. 1/1000, 1/10,000· Two National Samples of the Population of the United States, Description and Technical Documentation* Washington Government Printing Office, 1964

U.S., Department of Health, Education, and Welfare *Vital Statistics of the United States.* Vol 1 Washington Government Printing Office, 1968 (*a*)

———. Social Security Administration, Office of Research and Statistics *Workers under Social Security, 1960* Washington Government Printing Office, 1968 (*b*)

U.S , Office of Economic Opportunity, Day Care Policy Studies Group, Institute for Interdisciplinary Studies "Child Care Programs Estimation of Impacts and Evaluation of Alternative Federal Strategies " Report BOO-5121 Washington Office of Economic Opportunity, December 1971

Wachtel, Paul "The Returns to Investments in Higher Education Revisited " Mimeographed New York Nat Bur Econ Res , 1973

Wald, A "The Approximate Determination of Indifference Surfaces by Means of Engel Curves " *Econometrica* 8 (April 1940) 144–75

Wales, T "Estimation of a Labor Supply Curve for Self-employed Business Proprietors " *Internat Econ Rev* 14 (February 1973) 69–80

Weisbrod, Burton A , and Karpoff, Peter "Monetary Returns to College Education, Student Ability, and College Quality " *Rev Econ and Statis* 50 (November 1968) 491–97

Welch, Finis "Education in Production " *J P E* 78 (February 1970) 35–59.

Willis, Robert J "A New Approach to the Economic Theory of Fertility Behavior " *J P E* 81, pt. 2, suppl (March/April 1973) S14–64

Wilson, Frances *Labour in the South African Gold Mines, 1911–1969* Cambridge Cambridge Univ Press, 1972

Young, Kimball *Isn't One Wife Enough?* New York Holt, 1954

Zellner, Arnold "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias " *J American Statis Assoc* 57 (June 1962) 348–68

African Development

POPULAR GOVERNMENT IN AN AFRICAN TOWN

Kita, Mali

Nicholas S. Hopkins

Emphasizing process and idiom as well as structure, Nicholas S. Hopkins examines the strong local identity and massive popular participation that set the framework for political struggle in Kita, a small town in Western Mali Mr Hopkins shows that the long-term survival of an emerging national government depends upon the government's success in fostering economic and social development consonant with local values In Mali, the disparity between national and local notions of key political concepts adversely affected national sponsored development programs based on the cooperation of local participants

1972 *xxi, 245 pages* *Cloth $9.50*

DRAWERS OF WATER

Domestic Water Use in East Africa

Gilbert F. White, David J. Bradley, and Anne U. White

Based on several years of detailed field investigations in a variety of sites, this study considers systems of water use and their improvement in developing countries. Monetary and energy costs of obtaining water, health effects, productivity of water users, volume and allocation of water, and how water users choose among possible alternative sources of water are examined. The analyses of social costs and individual decision making will interest economists and anthropologists dealing with social process in developing areas.

1972 *328 pages* *illus.* *Cloth $14.00*

The University of Chicago Press

5801 South Ellis Avenue • Chicago 60637

On Modernization

ECONOMIC GROWTH IN CHINA AND INDIA, 1952-1970

A Comparative Appraisal

Subramanian Swamy

The governments that emerged in both China and India in the late 1940s struggled to transform their economies into modern industrial ones, each adopting an entirely different economic system Utilizing data from official records, Professor Swamy examines their success

Originally published as a supplement to *Economic Development and Cultural Change* journal

1973 *128 pages* *Cloth* $6 95,
Journal issue (*vol* *21, no 4, part 2*)

THE GOLDEN ROAD TO MODERNITY

Manning Nash

Considers the problem of transforming the newly independent, underdeveloped countries of Asia into modern ones by focusing on the structure and the nature of life in two villages in Upper Burma, near Mandalay The author provides a comprehensive picture of Burmese society by describing how the social and cultural heritage of communities affects the economic and political change inherent in modernization

1973 *344 pages* *Paper* $2 95

The
University of
Chicago
PRESS

Chicago

CHICAGO ESSAYS IN ECONOMIC DEVELOPMENT

Edited and with an Introduction by David Wall

In this collection of essays by members of the Chicago school of economics, David Wall shows that the School's methods go a long way toward the clarification and solution of the economic problems faced by the world's underdeveloped countries. This volume rectifies misunderstandings about the Chicago school and circulates more widely some of its best work. Contributors: Theodore Schultz, Harry G. Johnson, Arnold C. Harberger, Bert F. Hoselitz, Larry A. Sjaastad, and D. Gale Johnson

1972 *400 pages* *Cloth $14.00*

DUALISTIC ECONOMIC DEVELOPMENT

Theory and History

Allen C. Kelley, Jeffrey G. Williamson,
and Russell J. Cheetham

This innovative study provides an integrated application of formal economic theory, mathematical techniques of analysis, and economic development and structural change in the low-income economy. Illustrating the presentation is an examination of Japanese economic growth during the period 1885–1915

1972 *392 pages* *Cloth $12.50*

INVESTMENT IN EDUCATION

The Equity-Efficiency Quandary Workshop at the University of Chicago, June 7–10, 1971

Edited by Theodore W. Schultz

These workshop papers, first published as a supplement to the *Journal of Political Economy*, deal broadly with the unsettled social question of the effects of education upon the distribution of personal income.

1972 *298 pages* *Cloth $7.95*

Paper Journal of Political Economy (vol. 80, no. 3, part II) Institutions, $4.50, Individuals, $3.50; Twenty or more copies, $3.00 each

The University of Chicago Press
Chicago 60637